

Online Appendices

Deep-Learning-Based Causal Inference for Large-Scale Combinatorial Experiments

By Zikun Ye, Zhiqi Zhang, Dennis J. Zhang, Heng Zhang, Renyu Zhang

Appendix A: Technical Details

A.1. Proof of Theorem 1

Our proof is inspired by the analysis of universal approximation bounds of neural networks (e.g., [Kushilevitz and Mansour 1991](#)). At its core are the Fourier transformations of functions. Utilizing such transformations in our context, we develop a compact representation of the ATE as Boolean functions that allow us to connect to our *Generalized Sigmoid Form II*. The proof contains four steps. First, we give the Fourier representation of the function $f(\mathbf{t})$. Second, we rewrite the Fourier formula using indicator functions. Third, to remove the integral and obtain a finite population bound, we use a probabilistic method. Last, we approximate the indicator functions with our sigmoid link function.

Step 1 (Fourier representation). We first give the Fourier transformation of Boolean functions (e.g., [Kushilevitz and Mansour 1991](#)): For $f(\cdot) : \{0, 1\}^m \mapsto \mathbb{R}$, we always have

$$f(\mathbf{t}) = \sum_{\mathbf{w} \in \{0, 1\}^m} \hat{f}(\mathbf{w}) \chi_{\mathbf{w}}(\mathbf{t}), \quad \text{where } \hat{f}(\mathbf{w}) := \frac{1}{2^m} \sum_{\mathbf{t} \in \{0, 1\}^m} f(\mathbf{t}) \chi_{\mathbf{w}}(\mathbf{t}),$$

where

$$\chi_{\mathbf{w}}(\mathbf{t}) = \begin{cases} 1 & , \quad \text{if } \sum_{i=1}^m w_i t_i \bmod 2 = 0 \\ -1 & , \quad \text{if } \sum_{i=1}^m w_i t_i \bmod 2 = 1. \end{cases}$$

Note that this implies

$$f(\mathbf{t}) = \sum_{\mathbf{w} \in \{0, 1\}^m} \hat{f}(\mathbf{w}) \chi_{\mathbf{w}}(\mathbf{t}) = \sum_{\mathbf{w} \in \{0, 1\}^m} |\hat{f}(\mathbf{w})| \cos(\pi \mathbf{w}^\top \mathbf{t} + \delta(\mathbf{w}))$$

where $\delta(\mathbf{w}) = \pi \mathbb{1}\{\hat{f}(\mathbf{w}) < 0\}$.

Step 2 (Expansion with indicator functions). In the next, we expand the ATE, i.e., $f(\mathbf{t}) - f(\mathbf{0})$, using integrals and indicator functions.

$$\begin{aligned} & f(\mathbf{t}) - f(\mathbf{0}) \\ &= \sum_{\mathbf{w} \in \{0, 1\}^m} |\hat{f}(\mathbf{w})| \cos(\pi \mathbf{w}^\top \mathbf{t} + \delta(\mathbf{w})) - \cos(\delta(\mathbf{w})) \\ &= \sum_{\mathbf{w} \in \{0, 1\}^m} |\hat{f}(\mathbf{w})| \cdot \left[-\pi \int_0^{\mathbf{w}^\top \mathbf{t}} \sin(\pi z + \delta(\mathbf{w})) dz \right] \\ &= \sum_{\mathbf{w} \in \{0, 1\}^m} |\hat{f}(\mathbf{w})| \cdot \left[-\pi \int_0^{\sqrt{m} \|\mathbf{w}\|} \mathbb{1}(\mathbf{w}^\top \mathbf{t} - z \geq 0) \sin(\pi z + \delta(\mathbf{w})) dz + \pi \int_{-\sqrt{m} \|\mathbf{w}\|}^0 \mathbb{1}(\mathbf{w}^\top \mathbf{t} - z \leq 0) \sin(\pi z + \delta(\mathbf{w})) dz \right], \end{aligned} \tag{7}$$

where the last equality follows from the elementary inequality $|\mathbf{w}^\top \mathbf{t}| \leq \sqrt{m} \|\mathbf{w}\|$.

Step 3 (Sampling). Define $C_{f,m} := \sum_{\mathbf{w} \in \{0,1\}^m} |\hat{f}(\mathbf{w})| \sqrt{m} \|\mathbf{w}\|$. We use random sampling to remove the integrals in the ATE representation. Particularly, for each $\ell = 1, \dots, K$, we sample $(\mathbf{w}_\ell, z_\ell) \in \mathbb{R}^m \times \mathbb{R}$ in an *i.i.d* fashion as follows:

$$(\mathbf{w}_\ell, z_\ell) = \begin{cases} (0, 0), & \text{with probability } \frac{1}{3}, \\ \{0, 1\}^m \times (0, m\|\mathbf{w}\|], & \text{with probability density } \frac{|\hat{f}(\mathbf{w})|}{3C_{f,m}}, \\ \{0, 1\}^m \times [-m\|\mathbf{w}\|, 0), & \text{with probability density } \frac{|\hat{f}(\mathbf{w})|}{3C_{f,m}}. \end{cases}$$

Also, define the function

$$g(\mathbf{w}, z; \mathbf{t}) := \begin{cases} 3f(0)\mathbb{1}(\mathbf{0}^\top \mathbf{t} + 0 \geq 0), & \text{if } (\mathbf{w}, z) = \mathbf{0}, \\ -3\pi C_{f,m} \sin(\pi z + \delta(\mathbf{w}))\mathbb{1}(\mathbf{w}^\top \mathbf{t} - z \geq 0), & \text{if } (\mathbf{w}, z) \in \{0, 1\}^m \times (\mathbf{0}, \sqrt{m}\|\mathbf{w}\|], \\ 3\pi C_{f,m} \sin(\pi z + \delta(\mathbf{w}))\mathbb{1}(-\mathbf{w}^\top \mathbf{t} + z \geq 0), & \text{if } (\mathbf{w}, z) \in \{0, 1\}^m \times [-\sqrt{m}\|\mathbf{w}\|, \mathbf{0}). \end{cases}$$

Note that in the above we write $f(0) = f(0)\mathbb{1}(\mathbf{0}^\top \mathbf{t} + 0 \geq 0)$ for reasons that will be clear very soon. In view of (7), we notice that for any $\ell = 1, \dots, K$ and $\mathbf{t} \in \{0, 1\}^m$, $\mathbb{E}_{(\mathbf{w}_\ell, z_\ell)}[g(\mathbf{w}_\ell, z_\ell; \mathbf{t})] = f(\mathbf{t})$. This implies that

$$\begin{aligned} \mathbb{E}_{\{(\mathbf{w}_\ell, z_\ell)\}_{\ell=1}^K} \left[\sum_{\mathbf{t} \in \{0, 1\}^m} \left(f(\mathbf{t}) - \sum_{\ell=1}^K p(\ell)g(\mathbf{w}_\ell, z_\ell; \mathbf{t}) \right)^2 \right] &= \frac{1}{K} \sum_{\mathbf{t} \in \{0, 1\}^m} \text{Var}_{(\mathbf{w}_1, z_1)}[g(\mathbf{w}_1, z_1; \mathbf{t})] \\ &\leq \frac{1}{K} \sum_{\mathbf{t} \in \{0, 1\}^m} \mathbb{E}_{(\mathbf{w}_1, z_1)}[g(\mathbf{w}_1, z_1; \mathbf{t})^2] \leq \frac{9 \cdot 2^m}{K} f^2(0) \vee C_{f,m}^2, \end{aligned}$$

where the first equality and the first inequality follow from straightforward computation, and the second inequality follows from the definition of $g(\cdot, \cdot; \cdot)$. Therefore, there exists $\{(\tilde{\mathbf{w}}_\ell, \tilde{z}_\ell) : \ell = 1, \dots, K\}$ where each $(\tilde{\mathbf{w}}_\ell, \tilde{z}_\ell)$ is in the support of the random sampling distribution defined above such that

$$\frac{1}{2^m} \sum_{\mathbf{t} \in \{0, 1\}^m} \left(f(\mathbf{t}) - \frac{1}{K} \sum_{\ell=1}^K g(\tilde{\mathbf{w}}_\ell, \tilde{z}_\ell; \mathbf{t}) \right)^2 \leq \frac{9}{K} f^2(0) \vee C_{f,m}^2.$$

Step 4 (Approximation with sigmoid functions). Note that regardless of the value of (\mathbf{w}, z) , $g(\mathbf{w}, z; t)$ can be written in the form $g_1(\mathbf{w}, z)\mathbb{1}(\boldsymbol{\alpha}(\mathbf{w})^\top t + \beta(z) \geq 0)$, where $g_1(\mathbf{w}, z)$, $\boldsymbol{\alpha}(\mathbf{w})$ and $\beta(z)$ are some functions of \mathbf{w} and z . Further, such functions as $g_1(\mathbf{w}, z)\mathbb{1}(\boldsymbol{\alpha}(\mathbf{w})^\top t + \beta(z) \geq 0)$ can be well-approximated by scaled sigmoid functions. Indeed, $\frac{1}{1+e^{-\gamma z}} \rightarrow \mathbb{1}(z \geq 0)$ pointwise except for $z = 0$, when $\gamma \rightarrow \infty$. Thus, for any $\ell = 1, \dots, m$,

- if $(\tilde{\mathbf{w}}_\ell, \tilde{z}_\ell) = \mathbf{0}$, set $\theta_0(\mathbf{x}_\ell) = \gamma$, $\theta_1(\mathbf{x}_\ell) = \dots = \theta_m(\mathbf{x}_\ell) = 0$, and $\theta_{m+1}(\mathbf{x}_\ell) = 3Kp(\ell)f(0)$;
- if $(\tilde{\mathbf{w}}_\ell, \tilde{z}_\ell) \in \{0, 1\}^m \times (0, \sqrt{m}\|\tilde{\mathbf{w}}\|]$, set $\theta_0(\mathbf{x}_\ell) = -\gamma\tilde{z}_\ell$, $\theta_j(\mathbf{x}_\ell) = \gamma\tilde{w}_{\ell,j}$ for $j = 1, \dots, m$, and $\theta_{m+1}(\mathbf{x}_\ell) = -3\pi Kp(\ell)C_{f,m} \sin(\pi\tilde{z}_\ell + \delta(\tilde{\mathbf{w}}_\ell))$;
- and if $(\tilde{\mathbf{w}}_\ell, \tilde{z}_\ell) \in \{0, 1\}^m \times [-\sqrt{m}\|\tilde{\mathbf{w}}\|, 0)$, set $\theta_0(\tilde{\mathbf{w}}_\ell) = \gamma\tilde{z}_\ell$, $\theta_j(\mathbf{x}_\ell) = -\gamma\tilde{w}_{\ell,j}$ for $j = 1, \dots, m$, and $\theta_{m+1}(\mathbf{x}_\ell) = 3\pi Kp(\ell)C_{f,m} \sin(\pi\tilde{z}_\ell + \delta(\tilde{\mathbf{w}}_\ell))$.

Here we set γ large enough so that

$$\left| p(\ell)G(\boldsymbol{\theta}(\mathbf{x}_\ell), \mathbf{t}) - \frac{1}{K}g(\tilde{\mathbf{w}}_\ell, \tilde{z}_\ell; \mathbf{t}) \right| \leq \frac{1}{K},$$

uniformly over all $\ell = 1, \dots, K$ and $\mathbf{t} \in \{0, 1\}^m$. Thus,

$$\begin{aligned} &\frac{1}{2^m} \sum_{\mathbf{t} \in \{0, 1\}^m} \left(f(\mathbf{t}) - \sum_{\ell=1}^K p(\ell)G(\boldsymbol{\theta}(\mathbf{x}_\ell), \mathbf{t}) \right)^2 \\ &\leq \frac{2}{2^m} \sum_{\mathbf{t} \in \{0, 1\}^m} \left(f(\mathbf{t}) - \frac{1}{K} \sum_{\ell=1}^K g(\tilde{\mathbf{w}}_\ell, \tilde{z}_\ell; \mathbf{t}) \right)^2 + \frac{2K}{2^m} \sum_{\mathbf{t} \in \{0, 1\}^m} \sum_{\ell=1}^K \left(\frac{1}{K}g(\tilde{\mathbf{w}}_\ell, \tilde{z}_\ell; \mathbf{t}) - p(\ell)G(\boldsymbol{\theta}(\mathbf{x}_\ell), \mathbf{t}) \right)^2 \\ &\lesssim \frac{1}{K} (f^2(0) \vee C_{f,m}^2 + 1) \lesssim \frac{1}{K}, \end{aligned}$$

where the first inequality follows from repeated applications of the Young's inequality. We conclude the proof. \square

A.2. Regularity Assumptions

We make the following regularity assumption throughout the theoretical analysis of this paper.

- ASSUMPTION 3.** (a). $\mathbf{z}_i = (y_i, \mathbf{x}'_i, \check{\mathbf{t}}'_i)'$, $1 \leq i \leq n$, are i.i.d. copies from the population random variables $\mathbf{Z} = (Y, \mathbf{X}', \mathbf{T}')' \in \mathcal{Y} \times [-1, 1]^{d_{\mathbf{X}}} \times \{0, 1\}^m$, where \mathcal{Y} is the bounded support of the outcome Y .
(b). The parameter function $\boldsymbol{\theta}^*(\mathbf{x})$ is uniformly bounded. Furthermore, $\theta_k^*(\mathbf{x}) \in W^{p,\infty}([-1, 1]^{d_{\mathbf{X}}})$, $k = 1, 2, \dots, d_{\boldsymbol{\theta}}$, where for positive integers p , define the Sobolev ball $W^{p,\infty}([-1, 1]^{d_{\mathbf{X}}})$ of functions $h : \mathbb{R}^{d_{\mathbf{X}}} \mapsto \mathbb{R}$ with smoothness $p \in \mathbb{N}_+$ as,

$$W^{p,\infty}([-1, 1]^{d_{\mathbf{X}}}) := \left\{ h : \max_{\mathbf{r}, |\mathbf{r}| \leq p} \operatorname{ess\,sup}_{\mathbf{v} \in [-1, 1]^{d_{\mathbf{X}}}} |D^{\mathbf{r}} h(\mathbf{v})| \leq 1 \right\},$$

where $\mathbf{r} = (r_1, \dots, r_{d_{\mathbf{X}}})$, $|\mathbf{r}| = r_1 + \dots + r_{d_{\mathbf{X}}}$ and $D^{\mathbf{r}} h$ is the weak derivative.

We remark that Assumption 3(a) implies that the DGP is bounded, whereas Assumption 3(b) ensures that the ground-truth parameter functions are uniformly bounded, and enjoy sufficient smoothness so they can be accurately approximated by DNNs. The smoothness assumptions (see, also, Assumption 2 in Farrell et al. 2020) are critical to deriving the sufficiently fast convergence rate of the estimator $\hat{\boldsymbol{\theta}}(\cdot)$.

We also make the following assumption throughout our analysis to ensure the identifiability and sufficient convergence rate of our model. Let $\mathbf{t}(S) = (t_1(S), \dots, t_m(S))' \in \{0, 1\}^m$ denote the treatment assignment such that $t_i(S) = \mathbf{1}\{i \in S\}$, where $S \subset \{1, 2, \dots, m\}$, and define $\tilde{\mathbf{T}} := (1, \mathbf{T}')'$ and $\tilde{\mathbf{t}} := (1, \mathbf{t}')'$. Let $\lambda_{\min}(\cdot)$ denote the minimum eigenvalue of a symmetric matrix

ASSUMPTION 4. Any of the following conditions hold:

- (a) $G(\cdot, \cdot)$ is of the Multiplicative Form, $\lambda_{\min}(\mathbb{E}[\tilde{\mathbf{T}}\tilde{\mathbf{T}}'|\mathbf{X}])$ and $|\theta_0^*(\mathbf{X})|$ are uniformly bounded away from zero;
- (b) $G(\cdot, \cdot)$ is of the Standard Sigmoid Form, $\lambda_{\min}(\mathbb{E}[\tilde{\mathbf{T}}\tilde{\mathbf{T}}'|\mathbf{X}])$ is uniformly bounded away from zero;
- (c) $G(\cdot, \cdot)$ is of the Generalized Sigmoid Form I, $\lambda_{\min}(\mathbb{E}[\mathbf{T}\mathbf{T}'|\mathbf{X}])$, $|\theta_{m+1}^*(\mathbf{X})|$, and $\nu(\mathbf{t}(\emptyset)|\mathbf{X})$ are uniformly bounded away from zero;
- (d) $G(\cdot, \cdot)$ is of the Generalized Sigmoid Form II, $\lambda_{\min}(\mathbb{E}[\tilde{\mathbf{T}}\tilde{\mathbf{T}}'|\mathbf{X}])$ and $|\theta_{m+1}^*(\mathbf{x})|$ are uniformly bounded away from zero, and there exists a triplet (i, S_1, S_2) ($i \in \{1, \dots, m\}$, $S_1, S_2 \subset \{1, 2, \dots, m\}$) such that $i \notin S_1$, $i \notin S_2$, $S_1 \neq S_2$, and $\nu(\mathbf{t}(S_1 \cup \{i\})|\mathbf{X}) \cdot \nu(\mathbf{t}(S_1)|\mathbf{X}) \cdot \nu(\mathbf{t}(S_2 \cup \{i\})|\mathbf{X}) \cdot \nu(\mathbf{t}(S_2)|\mathbf{X})$ and $|G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t}(S_1 \cup \{i\}))G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t}(S_2)) - G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t}(S_2 \cup \{i\}))G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t}(S_1))|$ are both uniformly bounded away from zero.

Technically, we require that these conditions are satisfied uniformly on a set with probability one. For simplicity, we skip almost everywhere in these statements. Also, the assumptions $\lambda_{\min}(\mathbb{E}[\tilde{\mathbf{T}}\tilde{\mathbf{T}}'|\mathbf{X}])$ or $\lambda_{\min}(\mathbb{E}[\mathbf{T}\mathbf{T}'|\mathbf{X}])$ being uniformly bounded above zero is satisfied as long as all m individual treatment conditions and the full control condition are assigned with positive probability, i.e., $\nu((0, 0, \dots, 0)'|\mathbf{x}) > c$ and $\nu((1, 0, \dots, 0)'|\mathbf{x})$, $\nu((0, 1, \dots, 0)'|\mathbf{x})$, ..., $\nu((0, 0, \dots, 1)'|\mathbf{x}) > c$, for some $c > 0$ almost everywhere, which is a fairly mild assumption. Therefore, these conditions state that $\Omega(m)$ treatment assignments are necessary.

At a high level, we indeed can extend our framework to cases with higher-order interactions between individual treatments, and Theorem 2 remains valid. However, the number of observable treatment combinations must increase. For instance, to ensure the valid identification and inference for the model with quadratic interactions mentioned above, one will need $\Omega(m^2)$ treatment combinations assigned with a positive probability. Following the same inference framework, we notice that the key is to guarantee the identifiability of the model and the sufficient curvature condition in the presence of higher-order interactions.

Let us first consider the case where all second-order treatment interactions are included in the link function. In this regard, upon inspection of the proof for the *Multiplicative Form*, *Standard Sigmoid Form*, and *Generalized Sigmoid Form I*, we note that only mild modifications of the assumptions on the data generation process ensure the validity of our results. More specifically, one only needs to redefine $\tilde{\mathbf{T}} = (1, \mathbf{T}', \bar{\mathbf{T}}')'$ where $\bar{\mathbf{T}} = (T_i T_j : \forall i < j, 1 \leq i, j \leq m)$. That is, we extend the definition of $\tilde{\mathbf{T}}$ to include all quadratic interactions. Then, all analyses of the *Multiplicative Form* and *Standard Sigmoid Form* go through unchanged. Particularly, we notice that in this case $\tilde{\mathbf{T}}$ is of dimension $\Theta(m^2)$, so to ensure that the minimum eigenvalue of $\mathbb{E}[\tilde{\mathbf{T}}\tilde{\mathbf{T}}'|\mathbf{X}]$ is uniformly bounded from below above zero, the support of treatment assignment distribution should have cardinality of order $\Omega(m^2)$. This is in sharp contrast with our results with linear terms only, which only requires $\Omega(m)$ treatment combinations with positive probability, highlighting the price for modeling the higher-order treatment interactions and estimating the associated nuisance parameters.

With the *Generalized Sigmoid Form II*, the analysis is more involved. Consider the following link function

$$G(\boldsymbol{\theta}(\mathbf{x}), \mathbf{t}) = \frac{\theta_{m+1}(\mathbf{x})}{1 + \exp(-(\theta_0(\mathbf{x}) + \sum_{i=1}^m \theta_i(\mathbf{x}) t_i + \sum_{i < j} \theta_{ij}(\mathbf{x}) t_i t_j))}.$$

As discussed above, we need the minimum eigenvalue of $\mathbb{E}[\tilde{\mathbf{T}}\tilde{\mathbf{T}}'|\mathbf{X}]$ is uniformly bounded from below above zero. Moreover, to estimate $\theta_{m+1}(\mathbf{x})$, fix any i, j with $1 \leq i < j \leq m$ and consider S and \tilde{S} such that (1) $i, j \notin S, \tilde{S}$, (2) $S \neq \tilde{S}$, and (3) $\prod_{k=1}^8 \nu(\mathbf{t}(S_k)|\mathbf{X}) \geq \tilde{c}_1 > 0$ for some positive \tilde{c}_1 , where $S_1 = S$, $S_2 = S \cup \{i\}$, $S_3 = S \cup \{j\}$, $S_4 = S \cup \{i, j\}$, $S_5 = \tilde{S}$, $S_6 = \tilde{S} \cup \{i\}$, $S_7 = \tilde{S} \cup \{j\}$, $S_8 = \tilde{S} \cup \{i, j\}$. Then it follows that

$$\frac{\left(\frac{\theta_{m+1}(\mathbf{X})}{G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_1))} - 1\right) \left(\frac{\theta_{m+1}(\mathbf{X})}{G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_4))} - 1\right)}{\left(\frac{\theta_{m+1}(\mathbf{X})}{G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_2))} - 1\right) \left(\frac{\theta_{m+1}(\mathbf{X})}{G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_3))} - 1\right)} = \exp(-(\theta_i(\mathbf{X}) + \theta_j(\mathbf{X}) + \theta_{ij}(\mathbf{X}))) \quad (8)$$

and

$$\frac{\left(\frac{\theta_{m+1}(\mathbf{X})}{G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_5))} - 1\right) \left(\frac{\theta_{m+1}(\mathbf{X})}{G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_8))} - 1\right)}{\left(\frac{\theta_{m+1}(\mathbf{X})}{G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_6))} - 1\right) \left(\frac{\theta_{m+1}(\mathbf{X})}{G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_7))} - 1\right)} = \exp(-(\theta_i(\mathbf{X}) + \theta_j(\mathbf{X}) + \theta_{ij}(\mathbf{X}))). \quad (9)$$

Under the assumption that $\theta_{m+1}(\mathbf{X}) \neq 0$, putting Eqns (8) and (9) together allows us to cancel out the term $\exp(-(\theta_i(\mathbf{X}) + \theta_j(\mathbf{X}) + \theta_{ij}(\mathbf{X})))$ and build a cubic equation of $\theta_{m+1}(\mathbf{X})$ with the products of $G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_i))$ ($i = 1, \dots, 8$) as the coefficients. This cubic equation admits at most 3 real solutions. As long as there are 3 such (i, j) pairs (and the correspondingly sets S and \tilde{S}), we can uniquely identify $\theta_{m+1}(\mathbf{X})$. Therefore, we need $\Omega(m^2)$ treatment combinations with positive probability in the treatment assignment mechanism to ensure identification.

The argument above aligns with our analysis of the identification condition for *Generalized Sigmoid Form II* detailed in Appendix A.3, yet it becomes significantly more complex in the presence of second-order treatment

interactions. More generally, with higher-order interactions, we can establish equations similar to (8) and (9) to formulate primitive conditions for identifying $\theta_{m+1}(\mathbf{X})$, which makes the identification condition more restrictive and practically infeasible. Consequently, carefully trading off the generality, tractability, and practicality of our framework, we have decided to focus on the *Generalized Sigmoid Form II* link function.

A.3. Influence Function

The following lemma from Farrell et al. (2020) formally states a generic result regarding the influence function, which proves useful to derive the influence function in our setting (Proposition 2). The proof follows the pathwise derivative method originated in Newey (1994).

LEMMA 1 (Theorem 2 in Farrell et al. (2020)). *For all $\mathbf{t} \in \{0, 1\}^m$, suppose the following conditions hold uniformly in the given conditioning elements. (i) (1) holds and identifies $\boldsymbol{\theta}^*(\cdot)$. (ii) $\mathbb{E}[\ell_{\boldsymbol{\theta}}(Y, \check{\mathbf{t}}, \boldsymbol{\theta}^*(\mathbf{x})) | \mathbf{X} = \mathbf{x}, \mathbf{T} = \check{\mathbf{t}}] = 0$. (iii) $\boldsymbol{\Lambda}(\mathbf{x}) := \mathbb{E}[\ell_{\boldsymbol{\theta}\boldsymbol{\theta}}(Y, \mathbf{T}, \boldsymbol{\theta}(\mathbf{x})) | \mathbf{X} = \mathbf{x}]$ is invertible with uniformly bounded inverse. (iv) Parameter $\mu(\mathbf{t})$ is identified, pathwise differentiable, and H and ℓ are thrice continuously differentiable in $\boldsymbol{\theta}$. (v) $H(\mathbf{X}, \boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t}, \mathbf{t}_0)$ and $\ell_{\boldsymbol{\theta}}(Y, \mathbf{T}, \boldsymbol{\theta}^*(\mathbf{X}))$ possess $q > 4$ finite absolute moments and positive variance. Then for the treatment effect $\mu(\mathbf{t})$, the Neyman orthogonal score is $\psi(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\Lambda}; \mathbf{t}, \mathbf{t}_0) - \mu(\mathbf{t})$, where*

$$\psi(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\Lambda}; \mathbf{t}, \mathbf{t}_0) = H(\mathbf{x}, \boldsymbol{\theta}(\mathbf{x}); \mathbf{t}, \mathbf{t}_0) - H_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\theta}(\mathbf{x}); \mathbf{t}, \mathbf{t}_0)' \boldsymbol{\Lambda}(\mathbf{x})^{-1} \ell_{\boldsymbol{\theta}}(y, \check{\mathbf{t}}, \boldsymbol{\theta}(\mathbf{x})), \quad (10)$$

where $\ell_{\boldsymbol{\theta}}, H_{\boldsymbol{\theta}}$ are $d_{\boldsymbol{\theta}}$ -dimensional vectors of first order derivatives, and $\ell_{\boldsymbol{\theta}\boldsymbol{\theta}}$ is the $d_{\boldsymbol{\theta}} \times d_{\boldsymbol{\theta}}$ Hessian matrix of ℓ , with $\{k_1, k_2\}$ element defined by $\partial^2 \ell / \partial \theta_{k_1} \partial \theta_{k_2}$.

A.4. Proof of Proposition 1

We first present without proof a key convergence result inherited from Farrell et al. (2020).

LEMMA 2 (Theorem 1 in Farrell et al. (2020)). *Suppose Assumption 3, and the following regularity assumptions hold,*

- (a). (NONPARAMETRIC IDENTIFIABILITY) *The parameter function $\boldsymbol{\theta}^*(\mathbf{x})$ can be nonparametrically identified in DGP (1).*
- (b). (LIPSCHITZ CONTINUITY) *There exists a positive constant C_{ℓ} such that, for any $\boldsymbol{\theta}(\cdot), \tilde{\boldsymbol{\theta}}(\cdot)$ and \mathbf{x} ,*

$$|\ell(y, \mathbf{t}, \boldsymbol{\theta}(\mathbf{x})) - \ell(y, \mathbf{t}, \tilde{\boldsymbol{\theta}}(\mathbf{x}))| \leq C_{\ell} \|\boldsymbol{\theta}(\mathbf{x}) - \tilde{\boldsymbol{\theta}}(\mathbf{x})\|_2, \quad (11)$$

- (c). (SUFFICIENT CURVATURE) *There exist positive constants c_1 and c_2 such that, for any $\boldsymbol{\theta}(\cdot) \in \mathcal{F}_{DNN}$,*

$$c_1 \mathbb{E}[\|\boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta}^*(\mathbf{X})\|_2^2] \leq \mathbb{E}[\ell(Y, \mathbf{T}, \boldsymbol{\theta}(\mathbf{X}))] - \mathbb{E}[\ell(Y, \mathbf{T}, \boldsymbol{\theta}^*(\mathbf{X}))] \leq c_2 \mathbb{E}[\|\boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta}^*(\mathbf{X})\|_2^2]. \quad (12)$$

With the structured DNN of width $H = O(n^{\frac{d_X}{2(p+d_X)}} \log^2 n)$ and depth $L = O(\log n)$ as illustrated in Figure 3, there exists a constant C such that

$$\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\|_{L_2(\mathbf{X})}^2 \lesssim n^{-\frac{p}{p+d_X}} \log^8 n + \frac{\log \log n}{n}$$

and

$$\mathbb{E}_n[(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)^2] \lesssim n^{-\frac{p}{p+d_X}} \log^8 n + \frac{\log \log n}{n}$$

for n large enough with probability at least $1 - \exp(n^{-\frac{d_X}{p+d_X}} \log^8 n)$, for $k = 1, \dots, d_{\boldsymbol{\theta}}$.

Assumptions in Lemma 2 are natural and common in the nonparametric M-estimation literature, which consists of three parts: (a) the nonparametric identifiability of $\theta^*(\cdot)$, (b) the Lipschitz continuity of loss (i.e., (11)), and (c) the sufficient curvature of loss (i.e., (12)). Whereas the Lipschitz continuity condition is mild and easy to check in our setting, the identifiability of $\theta^*(\cdot)$ and the sufficient curvature condition are non-trivial and should be verified carefully. In particular, the sufficient curvature condition (12) is usually implied by a proper choice of the link function $G(\cdot, \cdot)$ and the treatment assignment mechanism $\nu(\cdot | \cdot)$. This condition helps translate the convergence of outcomes Y into that of parameter functions $\theta(\cdot)$.

To obtain the convergence results in Proposition 1, it suffices to verify the assumptions in Lemma 2 are satisfied. The Lipschitz condition in Assumption (b) can be easily satisfied by for all our proposed link functions in Section 3.3 because all G functions are sufficiently smooth with bounded \mathbf{X} , \mathbf{T} , and θ . Since the square loss function and our link functions are differentiable, the second inequality in the curvature condition is satisfied for all of our link functions. In particular, we note the identity

$$\mathbb{E}[\ell(Y, \mathbf{T}, \theta(\mathbf{X}))] - \mathbb{E}[\ell(Y, \mathbf{T}, \theta^*(\mathbf{X}))] = \mathbb{E}\left[(G(\theta(\mathbf{X}), \mathbf{T}) - G(\theta^*(\mathbf{X}), \mathbf{T}))^2\right] \quad (13)$$

by DGP (1), which then implies by the mean value theorem

$$\mathbb{E}[\ell(Y, \mathbf{T}, \theta(\mathbf{X}))] - \mathbb{E}[\ell(Y, \mathbf{T}, \theta^*(\mathbf{X}))] = \mathbb{E}[(\theta(\mathbf{X}) - \theta^*(\mathbf{X}))' G_\theta(\tilde{\theta}(\mathbf{X}), \mathbf{T}) G_\theta(\tilde{\theta}(\mathbf{X}), \mathbf{T})' (\theta(\mathbf{X}) - \theta^*(\mathbf{X}))], \quad (14)$$

where $\tilde{\theta}(\mathbf{X})$ is such that $\tilde{\theta}_i(\mathbf{X}) \in [\theta_i^*(\mathbf{X}), \theta_i(\mathbf{X})]$ for all component i . Since all variables are bounded, $G_\theta(\tilde{\theta}(\mathbf{X}), \mathbf{T}) G_\theta(\tilde{\theta}(\mathbf{X}), \mathbf{T})'$ is also uniformly bounded and the claim follows.

Consequently, it suffices to verify the nonparametric identifiability and the first inequality in the curvature condition in Lemma 2 for different forms of G functions to guarantee the convergence of structured DNNs. To simplify the notation, we define the one-dimension sigmoid function as $S(x) := 1/(1 + \exp(-x))$. In the following, we prove that, for each proposed link functions in Assumption 1, the conditions of Lemma 2 hold. We also remark that the constants may be different for different forms of the link function.

Standard Sigmoid Form. In this part, we start with the standard sigmoid form,

$$G(\theta(\mathbf{x}), \mathbf{t}) := \frac{a}{1 + \exp\left(-(\theta_0(\mathbf{x}) + \theta_1(\mathbf{x})t_1 + \dots + \theta_m(\mathbf{x})t_m)\right)} + b, \quad (15)$$

where the constants $a \neq 0$, b are known.

Proof. Because the sigmoid function is invertible, it suffices to verify the identifiability and sufficient curvature conditions for the linear link function $\theta(\mathbf{x})'\tilde{\mathbf{t}}$. Suppose $\mathbb{E}[Y | \mathbf{X} = \mathbf{x}, \mathbf{T} = \mathbf{t}] = G(\hat{\theta}(\mathbf{x})'\tilde{\mathbf{t}}) = G(\theta^*(\mathbf{x})'\tilde{\mathbf{t}})$. Next, we show $\theta^*(\cdot)$ can be nonparametrically identified, i.e., $\hat{\theta}(\mathbf{X}) = \theta^*(\mathbf{X})$. For all $\tilde{\mathbf{t}}$, we have,

$$0 = \hat{\theta}(\mathbf{X})'\tilde{\mathbf{t}} - \theta^*(\mathbf{X})'\tilde{\mathbf{t}} = (\hat{\theta}(\mathbf{X}) - \theta^*(\mathbf{X}))'\tilde{\mathbf{t}},$$

which implies,

$$0 = \mathbb{E}[(\hat{\theta}(\mathbf{X}) - \theta^*(\mathbf{X}))'\tilde{\mathbf{t}}^2 | \mathbf{X}] = (\hat{\theta}(\mathbf{X}) - \theta^*(\mathbf{X}))'\mathbb{E}[\tilde{\mathbf{t}}\tilde{\mathbf{t}}' | \mathbf{X}] (\hat{\theta}(\mathbf{X}) - \theta^*(\mathbf{X})).$$

Because $\mathbb{E}[\tilde{\mathbf{t}}\tilde{\mathbf{t}}' | \mathbf{X}]$ is positive definite uniformly with respect to \mathbf{X} , it must hold that $\hat{\theta}(\mathbf{X}) - \theta^*(\mathbf{X}) = 0$ almost everywhere, which concludes the proof of identifiability.

We show that the first inequality in the sufficient curvature condition is satisfied. Similar to (14), it holds that

$$\mathbb{E}[\ell(Y, \mathbf{T}, \boldsymbol{\theta}(\mathbf{X}))] - \mathbb{E}[\ell(Y, \mathbf{T}, \boldsymbol{\theta}^*(\mathbf{X}))] = \mathbb{E}\left[((\boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta}^*(\mathbf{X}))' \tilde{\mathbf{T}})^2 \cdot S^2(w(\mathbf{X}, \tilde{\mathbf{T}})) (1 - S(w(\mathbf{X}, \tilde{\mathbf{T}})))^2\right] \quad (16)$$

$$\geq \tilde{c}_1 \mathbb{E}\left[((\boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta}^*(\mathbf{X}))' \tilde{\mathbf{T}})^2\right] \quad (17)$$

where $w(\mathbf{X}, \tilde{\mathbf{T}})$ is some function such that $w(\mathbf{X}, \tilde{\mathbf{T}}) \in [\boldsymbol{\theta}^*(\mathbf{X})' \tilde{\mathbf{T}}, \boldsymbol{\theta}(\mathbf{X})' \tilde{\mathbf{T}}]$ and the inequality follows since all of the variables are bounded. Then, under the assumption $\mathbb{E}[\tilde{\mathbf{T}} \tilde{\mathbf{T}}' | \mathbf{X}]$ is uniformly bounded away from zero and following the law of total expectation, we derive that,

$$\begin{aligned} \mathbb{E}_{\mathbf{X}, \mathbf{T}}[((\boldsymbol{\theta}(\mathbf{X})' \tilde{\mathbf{T}} - \boldsymbol{\theta}^*(\mathbf{X})' \tilde{\mathbf{T}})^2)] &= a^2 \mathbb{E}_{\mathbf{X}}[((\boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta}^*(\mathbf{X}))' \mathbb{E}[\tilde{\mathbf{T}} \tilde{\mathbf{T}}' | \mathbf{X}] (\boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta}^*(\mathbf{X})))] \\ &\geq \tilde{c}_2 \mathbb{E}_{\mathbf{X}}[(\boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta}^*(\mathbf{X}))' (\boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta}^*(\mathbf{X}))], \end{aligned}$$

for some $\tilde{c}_2 > 0$, which concludes the proof of sufficient curvature, i.e.,

$$\mathbb{E}[\ell(Y, \mathbf{T}, \boldsymbol{\theta}(\mathbf{X}))] - \mathbb{E}[\ell(Y, \mathbf{T}, \boldsymbol{\theta}^*(\mathbf{X}))] \geq \tilde{c}_1 \tilde{c}_2 \mathbb{E}_{\mathbf{X}}[(\boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta}^*(\mathbf{X}))' (\boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta}^*(\mathbf{X}))] \quad (18)$$

This concludes the proof. □

Multiplicative Form. Consider the link function

$$G(\boldsymbol{\theta}(\mathbf{x}), \mathbf{t}) = \theta_0(\mathbf{x})(1 + \theta_1(\mathbf{x})t_1) \dots (1 + \theta_m(\mathbf{x})t_m), \quad (19)$$

where $\mu \leq 1 + \theta_k(\mathbf{x}) \leq M$, $k = 1, \dots, m$, uniformly in \mathbf{x} by assumption.

Proof. For simplicity of the proof, let us assume throughout $\theta_0(\mathbf{x}) > 0$ for \mathbf{x} almost everywhere. Other cases can be proved similarly. These conditions guarantee that $\log(G(\boldsymbol{\theta}(\mathbf{x}), \mathbf{t}))$ is well-defined. The proof of nonparametric identifiability and the curvature conditions can be verified as shown in the following:

$$\begin{aligned} \log G &= \log \theta_0(\mathbf{x})(1 + \theta_1(\mathbf{x})t_1) \dots (1 + \theta_m(\mathbf{x})t_m) \\ &= \log \theta_0(\mathbf{x}) + \log(1 + \theta_1(\mathbf{x})t_1) + \dots + \log(1 + \theta_m(\mathbf{x})t_m) \\ &= \log \theta_0(\mathbf{x}) + \log(1 + \theta_1(\mathbf{x}))t_1 + \dots + \log(1 + \theta_m(\mathbf{x}))t_m, \end{aligned}$$

where the last equality follows from that $t_i = 0, 1$ for all $i = 1, 2, \dots, m$. Hence, the *Multiplicative Form* is equivalent to $G(\boldsymbol{\theta}(\mathbf{x}), \mathbf{t}) = \exp(a(\mathbf{x}) + b_1(\mathbf{x})t_1 + \dots + b_m(\mathbf{x})t_m)$, with $a(\mathbf{x}) = \log \theta_0(\mathbf{x})$, and $b_k(\mathbf{x}) = \log(1 + \theta_k(\mathbf{x}))$. Because the exponential function is monotone and smooth, to satisfy the identification and sufficient curvature condition, we only need that $\mathbb{E}[\tilde{\mathbf{T}} \tilde{\mathbf{T}}' | \mathbf{X} = \mathbf{x}]$ are invertible uniformly in \mathbf{x} , where $\tilde{\mathbf{T}}' = (1, \mathbf{T}')$, i.e., Assumption 4(a). The proof follows from the same argument as the *Standard Sigmoid Form*. □

Generalized Sigmoid Form I. Next, we consider

$$G(\boldsymbol{\theta}(\mathbf{x}), \mathbf{t}) := \frac{\theta_{m+1}(\mathbf{x})}{1 + \exp(-(\theta_1(\mathbf{x})t_1 + \dots + \theta_m(\mathbf{x})t_m))}, \quad (20)$$

where $\theta_{m+1}(\mathbf{x}) \in \mathbb{R}$ can capture the range of the expected outcome, and the sign of $\theta_i(\mathbf{x})$ ($i = 1, 2, \dots, m$) represents whether the experiment i has a positive or negative effect.

Proof. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_-(\mathbf{x})', \theta_{m+1}(\mathbf{x}))'$, where $\boldsymbol{\theta}_-(\mathbf{x}) = (\theta_1(\mathbf{x}), \dots, \theta_m(\mathbf{x}))'$. Hence, we rewrite $G(\boldsymbol{\theta}(\mathbf{x}), \mathbf{t}) = \theta_{m+1}(\mathbf{x})S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t})$.

First, we show that $\boldsymbol{\theta}^*(\cdot)$ can be nonparametrically identified. Assume that $\mathbb{E}[Y|\mathbf{X}, \mathbf{T}] = \hat{\theta}_{m+1}(\mathbf{x})S(\hat{\boldsymbol{\theta}}_-(\mathbf{X})'\mathbf{T}) = \theta_{m+1}^*(\mathbf{X})S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T})$. Because $\nu(\mathbf{0}|\mathbf{X}) = \mathbb{P}[\mathbf{T} = \mathbf{0}|\mathbf{X}] > 0$ uniformly and $S(0) = 0.5 > 0$, we have $\hat{\theta}_{m+1}(\mathbf{X}) = \theta_{m+1}^*(\mathbf{X})$. Because $\theta_{m+1}^*(\mathbf{X})S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}) \neq 0$ by assumption and $\hat{\theta}_{m+1}(\mathbf{X}) = \theta_{m+1}^*(\mathbf{X})$, we have $S(\hat{\boldsymbol{\theta}}_-(\mathbf{X})'\mathbf{T}) = S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T})$. Also, because $S(\cdot)$ is continuously invertible, following the similar argument for the *Standard Sigmoid Form*, we derive that

$$0 = \mathbb{E}[(\hat{\boldsymbol{\theta}}_-(\mathbf{X})'\mathbf{T} - \boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T})^2|\mathbf{X}] = (\hat{\boldsymbol{\theta}}_-(\mathbf{X}) - \boldsymbol{\theta}_-(\mathbf{X}))'\mathbb{E}[\mathbf{T}\mathbf{T}'|\mathbf{X}](\hat{\boldsymbol{\theta}}_-(\mathbf{X}) - \boldsymbol{\theta}_-(\mathbf{X})).$$

Since $\mathbb{E}[\mathbf{T}\mathbf{T}'|\mathbf{X}] \succ 0$, it follows that $\hat{\boldsymbol{\theta}}_-(\mathbf{X}) = \boldsymbol{\theta}_-(\mathbf{X})$. Together with $\hat{\theta}_{m+1}(\mathbf{x}) = \theta_{m+1}^*(\mathbf{x})$, we conclude the proof of identifiability.

Next, we show the sufficient curvature condition. Using (13), we obtain $\mathbb{E}[\ell(Y, \mathbf{T}, \boldsymbol{\theta}(\mathbf{X}))] - \mathbb{E}[\ell(Y, \mathbf{T}, \boldsymbol{\theta}^*(\mathbf{X}))] = \mathbb{E}[(\theta_{m+1}(\mathbf{X})S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}) - \theta_{m+1}^*(\mathbf{X})S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}))^2]$. Hence, the sufficient curvature condition is equivalent to that there exists a constant $c_1 > 0$ such that

$$c_1 \mathbb{E}[\sum_{i=1}^{m+1} (\theta_i(\mathbf{X}) - \theta_i^*(\mathbf{X}))^2] \leq \mathbb{E}[(\theta_{m+1}(\mathbf{X})S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}) - \theta_{m+1}^*(\mathbf{X})S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}))^2].$$

Since $\mathbb{P}[\mathbf{T} = \mathbf{0}|\mathbf{X}]$ is uniformly bounded away from zero, which implies that $|G(\boldsymbol{\theta}^*(\mathbf{X})'\mathbf{t}(\emptyset))| = |\theta_{m+1}^*(\mathbf{X})S(0)| = |\theta_{m+1}(\mathbf{X})|/2$ with probability uniformly bounded away from zero for all \mathbf{X} , there must exist a constant $\tilde{c}_1 > 0$ such that,

$$\mathbb{E}_{\mathbf{X}, \mathbf{T}} \left[(\theta_{m+1}(\mathbf{X})S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}) - \theta_{m+1}^*(\mathbf{X})S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}))^2 \right] \geq \tilde{c}_1 \mathbb{E}_{\mathbf{X}} \left[(\theta_{m+1}(\mathbf{X}) - \theta_{m+1}^*(\mathbf{X}))^2 \right]. \quad (21)$$

With the condition $\mathbb{E}[\mathbf{T}\mathbf{T}'|\mathbf{X}]$ uniformly bounded away from zero, using the argument similar to (18), we can show that there exists a constant $\tilde{c}_2 > 0$, the following inequality holds

$$\mathbb{E}_{\mathbf{X}, \mathbf{T}} [(S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}) - S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}))^2] \geq \tilde{c}_2 \mathbb{E}_{\mathbf{X}} \left[\sum_{i=1}^m (\theta_i(\mathbf{X}) - \theta_i^*(\mathbf{X}))^2 \right]. \quad (22)$$

Then, for any realized \mathbf{x} and \mathbf{t} , we have the following decomposition

$$\begin{aligned} & \left| \theta_{m+1}^*(\mathbf{x})S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t}) - \theta_{m+1}(\mathbf{x})S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t}) \right| \\ & \geq \left| \theta_{m+1}^*(\mathbf{x})S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t}) - \theta_{m+1}^*(\mathbf{x})S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t}) \right| - \left| \theta_{m+1}^*(\mathbf{x})S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t}) - \theta_{m+1}(\mathbf{x})S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t}) \right| \\ & = |\theta_{m+1}^*(\mathbf{x})| \cdot |S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t}) - S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t})| - |\theta_{m+1}^*(\mathbf{x}) - \theta_{m+1}(\mathbf{x})| \cdot S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t}). \end{aligned}$$

Since $|\theta_{m+1}^*(\mathbf{x})| > 0$, rearranging the terms implies that

$$|S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t}) - S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t})| \leq \frac{|\theta_{m+1}^*(\mathbf{x})S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t}) - \theta_{m+1}(\mathbf{x})S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t})| + |\theta_{m+1}^*(\mathbf{x}) - \theta_{m+1}(\mathbf{x})|S(\boldsymbol{\theta}_-(\mathbf{x})'\mathbf{t})}{|\theta_{m+1}^*(\mathbf{x})|}.$$

If $|A| \leq |B| + |C|$, one can get $|A|^2 \leq |B|^2 + |C|^2 + 2|BC| \leq 2|B|^2 + 2|C|^2$. Using this elementary identity, together with the uniform boundedness of $\theta_{m+1}^*(\mathbf{x})$, we arrive at that there exists some constant $\tilde{c}_3 > 0$ such that

$$\begin{aligned} \tilde{c}_3 \mathbb{E} \left[(S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}) - S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}))^2 \right] & \leq \mathbb{E} \left[(\theta_{m+1}(\mathbf{X})S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}) - \theta_{m+1}^*(\mathbf{X})S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}))^2 \right] \\ & \quad + \mathbb{E} \left[(\theta_{m+1}^*(\mathbf{X}) - \theta_{m+1}(\mathbf{X}))^2 \right] \\ & \leq \left(1 + \frac{1}{\tilde{c}_1} \right) \cdot \mathbb{E} [(\theta_{m+1}(\mathbf{X})S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}) - \theta_{m+1}^*(\mathbf{X})S(\boldsymbol{\theta}_-(\mathbf{X})'\mathbf{T}))^2] \\ & = \left(1 + \frac{1}{\tilde{c}_1} \right) \cdot \mathbb{E} [(G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{T}) - G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{T}))^2], \end{aligned} \quad (23)$$

where the second inequality follows from (21). Putting the inequalities (21), (22), and (23) together, we conclude that

$$c_1 \mathbb{E} \left[\|\boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta}^*(\mathbf{X})\|_2^2 \right] \leq \mathbb{E} \left[\left(G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{T}) - G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{T}) \right)^2 \right],$$

where $c_1 = \tilde{c}_1 \tilde{c}_2 \tilde{c}_3 / (1 + \tilde{c}_1 + \tilde{c}_2 \tilde{c}_3)$, which concludes the proof. \square

Generalized Sigmoid Form II. We consider

$$G(\boldsymbol{\theta}(\mathbf{x}), \mathbf{t}) := \frac{\theta_{m+1}(\mathbf{x})}{1 + \exp \left(-(\theta_0(\mathbf{x}) + \theta_1(\mathbf{x})t_1 + \cdots + \theta_m(\mathbf{x})t_m) \right)}. \quad (24)$$

Recall that $\mathbf{t}(S)$ represents the treatment assignment vector such that $t_i(S) = 1$ if and only if $i \in S$, where $S \subset \{1, 2, \dots, m\}$. We restate the following sufficient conditions of Assumption 4(d): Uniformly in \mathbf{X} ,

- the matrix $\mathbb{E}[\tilde{\mathbf{T}}\tilde{\mathbf{T}}'|\mathbf{X}] \succ 0$, where $\tilde{\mathbf{T}}' = (1, \mathbf{T}')$,
- $|\theta_{m+1}^*(\mathbf{X})| > 0$,
- and there exists a triplet $i \in \{1, \dots, m\}$, $S_1, S_2 \subset \{1, 2, \dots, m\}$ such that $i \notin S_1, i \notin S_2, S_1 \neq S_2$, and there exists a constant \tilde{c}_1, \tilde{c}_2 such that

$$\nu(\mathbf{t}(S_1 \cup \{i\})|\mathbf{X}) \cdot \nu(\mathbf{t}(S_1)|\mathbf{X}) \cdot \nu(\mathbf{t}(S_2 \cup \{i\})|\mathbf{X}) \cdot \nu(\mathbf{t}(S_2)|\mathbf{X}) \geq \tilde{c}_1 > 0,$$

and

$$|G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t}(S_1 \cup \{i\}))G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t}(S_2)) - G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t}(S_2 \cup \{i\}))G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t}(S_1))| \geq \tilde{c}_2 > 0. \quad (25)$$

We show that these conditions are sufficient to guarantee the identification and sufficient curvature near the ground truth, i.e., Proposition 1 holds for the *Generalized Sigmoid Form II*.

Proof. To flesh out the analysis, given each $\boldsymbol{\theta}(\cdot)$ and \mathbf{X} , let us define

$$g_1(\boldsymbol{\theta}, \mathbf{X}) = G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_1)), g_2(\boldsymbol{\theta}, \mathbf{X}) = G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_2)), g_3(\boldsymbol{\theta}, \mathbf{X}) = G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_1 \cup \{i\})),$$

$$g_4(\boldsymbol{\theta}, \mathbf{X}) = G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_2 \cup \{i\})), \text{ and } \mathbf{g}(\boldsymbol{\theta}, \mathbf{X}) = (g_1(\boldsymbol{\theta}, \mathbf{X}), g_2(\boldsymbol{\theta}, \mathbf{X}), g_3(\boldsymbol{\theta}, \mathbf{X}), g_4(\boldsymbol{\theta}, \mathbf{X}))'.$$

Then, note that for any fixed \mathbf{t}

$$\frac{\theta_{m+1}(\mathbf{X})}{G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t})} - 1 = \exp \left(-(\theta_0(\mathbf{X}) + \theta_1(\mathbf{X})t_1 + \cdots + \theta_m(\mathbf{X})t_m) \right).$$

As a result, we have

$$\begin{aligned} \frac{\theta_{m+1}(\mathbf{X})}{G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_1 \cup \{i\}))} - 1 &= \frac{\theta_{m+1}(\mathbf{X})}{G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_2 \cup \{i\}))} - 1 \\ \frac{\theta_{m+1}(\mathbf{X})}{G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_1))} - 1 &= \frac{\theta_{m+1}(\mathbf{X})}{G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_2))} - 1 \end{aligned} = \exp(-\theta_i(\mathbf{X})),$$

which implies that

$$\begin{aligned} \theta_{m+1}(\mathbf{X}) \cdot \left[\left(\frac{1}{g_3(\boldsymbol{\theta}, \mathbf{X})g_2(\boldsymbol{\theta}, \mathbf{X})} - \frac{1}{g_4(\boldsymbol{\theta}, \mathbf{X})g_1(\boldsymbol{\theta}, \mathbf{X})} \right) \cdot \theta_{m+1}(\mathbf{X}) \right. \\ \left. - \frac{1}{g_3(\boldsymbol{\theta}, \mathbf{X})} - \frac{1}{g_2(\boldsymbol{\theta}, \mathbf{X})} + \frac{1}{g_4(\boldsymbol{\theta}, \mathbf{X})} + \frac{1}{g_1(\boldsymbol{\theta}, \mathbf{X})} \right] = 0. \end{aligned}$$

By assumption, $\theta_{m+1}(\mathbf{X}) \neq 0$, it then holds that

$$\begin{aligned} & (g_4(\boldsymbol{\theta}, \mathbf{X})g_1(\boldsymbol{\theta}, \mathbf{X}) - g_3(\boldsymbol{\theta}, \mathbf{X})g_2(\boldsymbol{\theta}, \mathbf{X}))\theta_{m+1}(\mathbf{X}) \\ &= g_1(\boldsymbol{\theta}, \mathbf{X})g_2(\boldsymbol{\theta}, \mathbf{X})g_4(\boldsymbol{\theta}, \mathbf{X}) + g_1(\boldsymbol{\theta}, \mathbf{X})g_3(\boldsymbol{\theta}, \mathbf{X})g_4(\boldsymbol{\theta}, \mathbf{X}) - g_1(\boldsymbol{\theta}, \mathbf{X})g_2(\boldsymbol{\theta}, \mathbf{X})g_3(\boldsymbol{\theta}, \mathbf{X}) - g_2(\boldsymbol{\theta}, \mathbf{X})g_3(\boldsymbol{\theta}, \mathbf{X})g_4(\boldsymbol{\theta}, \mathbf{X}). \end{aligned} \quad (26)$$

Assume that there exist $\tilde{\boldsymbol{\theta}}(\mathbf{X})$ such that

$$G(\tilde{\boldsymbol{\theta}}(\mathbf{X}), \mathbf{T}) = G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{T}),$$

with the assignment mechanism $\nu(\cdot | \cdot)$. Under our assignment mechanism assumption that

$$\mathbb{P}\left[\mathbf{T} = \mathbf{t}(S_1 \cup \{i\}) \mid \mathbf{X}\right] \cdot \mathbb{P}\left[\mathbf{T} = \mathbf{t}(S_1) \mid \mathbf{X}\right] \cdot \mathbb{P}\left[\mathbf{T} = \mathbf{t}(S_2 \cup \{i\}) \mid \mathbf{X}\right] \cdot \mathbb{P}\left[\mathbf{T} = \mathbf{t}(S_2) \mid \mathbf{X}\right] > \tilde{c}_1 > 0, \quad (27)$$

for some $\tilde{c}_1 > 0$, we have that $G(\tilde{\boldsymbol{\theta}}(\mathbf{X}), \mathbf{t}) = G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t})$ for $\mathbf{t} = \mathbf{t}(S_1 \cup \{i\})$, $\mathbf{t}(S_1)$, $\mathbf{t}(S_2 \cup \{i\})$, $\mathbf{t}(S_2)$, i.e., $\mathbf{g}(\boldsymbol{\theta}, \mathbf{X}) = \mathbf{g}(\boldsymbol{\theta}^*, \mathbf{X})$. And clearly, $\tilde{\theta}_{m+1}(\mathbf{X}) \neq 0$, because $\tilde{\theta}_{m+1}^*(\mathbf{X}) \neq 0$. Then we have $\tilde{\theta}_{m+1}(\mathbf{X}) = \theta_{m+1}(\mathbf{X})$, which is implied by (26). The rest of the proof of identifiability is the same as the *Generalized Sigmoid Form I*.

Next, we move on to the first inequality of the sufficient curvature condition. Due to Assumption 3 and the bounded output of the neural network, there must exists a constant \tilde{c}_3 such that $G(\boldsymbol{\theta}(\mathbf{x}), \mathbf{t}) \leq \tilde{c}_3$ for some $\tilde{c}_3 > 0$ for all $\boldsymbol{\theta}(\cdot) \in \mathcal{F}_{\text{DNN}}$, \mathbf{X} and \mathbf{t} . Let

$$\mathcal{E} = \left\{ \mathbf{X} : |g_\ell(\boldsymbol{\theta}, \mathbf{X}) - g_\ell(\boldsymbol{\theta}^*, \mathbf{X})| \leq \epsilon \forall \ell = 1, \dots, 4, \text{ and } \theta_{m+1}(\mathbf{X}) \neq 0 \right\}$$

where $\epsilon = \tilde{c}_2/(8\tilde{c}_3)$. Let us assume from now on that $\mathbf{X} \in \mathcal{E}$. Define function $h(\mathbf{g}) = g_4g_1 - g_3g_2$. By assumption, we have $|h(\mathbf{g}(\boldsymbol{\theta}^*, \mathbf{X}))| \geq \tilde{c}_2$. Then, by mean value theorem, it must be that

$$h(\mathbf{g}(\boldsymbol{\theta}, \mathbf{X})) = h(\mathbf{g}(\boldsymbol{\theta}^*, \mathbf{X})) + \nabla h(\hat{\mathbf{g}}(\mathbf{X}))'(\mathbf{g}(\boldsymbol{\theta}, \mathbf{X}) - \mathbf{g}(\boldsymbol{\theta}^*, \mathbf{X})) \geq h(\mathbf{g}(\boldsymbol{\theta}^*, \mathbf{X})) - 4\tilde{c}_3\epsilon \geq h(\mathbf{g}(\boldsymbol{\theta}^*, \mathbf{X}))/2 \geq \tilde{c}_2/2 > 0, \quad (28)$$

where $\hat{\mathbf{g}}(\mathbf{X})$ is such that $\hat{g}_\ell(\mathbf{X}) \in [g_\ell(\boldsymbol{\theta}^*, \mathbf{X}), g_\ell(\boldsymbol{\theta}, \mathbf{X})]$ for $\ell = 1, \dots, 4$ and the first inequality holds because $|\hat{g}_\ell(\mathbf{X}) - g_\ell(\boldsymbol{\theta}^*, \mathbf{X})| \leq \epsilon$ for all $\ell = 1, \dots, 4$ and the second inequality holds due to the definition of ϵ . Therefore, if we define

$$f(\mathbf{g}) = (g_1g_2g_4 + g_1g_3g_4 - g_1g_2g_3 - g_1g_3g_4)/h(\mathbf{g}),$$

then given (26) it holds that $\theta_{m+1}(\mathbf{X}) = f(\mathbf{g}(\boldsymbol{\theta}, \mathbf{X}))$, which is well-defined since we have already shown that $h(\mathbf{g}(\boldsymbol{\theta}, \mathbf{X})) > 0$. Similarly, we have $\theta_{m+1}^*(\mathbf{X}) = f(\mathbf{g}(\boldsymbol{\theta}^*, \mathbf{X}))$. Then, we have

$$(\boldsymbol{\theta}_{m+1}(\mathbf{X}) - \boldsymbol{\theta}_{m+1}^*(\mathbf{X}))^2 = [\nabla f(\hat{\mathbf{g}}(\mathbf{X}))'(\mathbf{g}(\boldsymbol{\theta}, \mathbf{X}) - \mathbf{g}(\boldsymbol{\theta}^*, \mathbf{X}))]^2, \quad (29)$$

by the mean value theorem, where $\tilde{g}_\ell(\mathbf{X}) \in [g_\ell(\boldsymbol{\theta}^*, \mathbf{X}), g_\ell(\boldsymbol{\theta}, \mathbf{X})]$ for all $\ell = 1, \dots, 4$. Note that every term in $\nabla f(\hat{\mathbf{g}}(\mathbf{X}))$ is of the form $\text{poly}(\tilde{\mathbf{g}}(\mathbf{X}))/h^2(\tilde{\mathbf{g}}(\mathbf{X}))$, where $\text{poly}(\tilde{\mathbf{g}}(\mathbf{X}))$ is a polynomial of $\tilde{\mathbf{g}}(\mathbf{X})$ and uniformly bounded by assumption. By exactly the same argument to (28), we can still show that $h(\tilde{\mathbf{g}}(\mathbf{X})) > \tilde{c}_2/2$. Therefore, expanding the quadratic term on the right-hand side of (29), we obtain that for some constant \tilde{c}_3 ,

$$(\boldsymbol{\theta}_{m+1}(\mathbf{X}) - \boldsymbol{\theta}_{m+1}^*(\mathbf{X}))^2 \leq \tilde{c}_4 \cdot \|\mathbf{g}(\boldsymbol{\theta}, \mathbf{X}) - \mathbf{g}(\boldsymbol{\theta}^*, \mathbf{X})\|^2. \quad (30)$$

If $\mathbf{X} \in \mathcal{E}_1 = \left\{ \mathbf{X} : \theta_{m+1}(\mathbf{X}) = 0 \right\}$, it must be that $g(\boldsymbol{\theta}, \mathbf{X}) = 0$ by definition. Also, by the uniform boundedness assumption, there exists constant \tilde{c}_5 such that $g_\ell(\boldsymbol{\theta}^*, \mathbf{X}) \geq \tilde{c}_5$. Therefore, it must be that there is constant \tilde{c}_6 such that

$$(\boldsymbol{\theta}_{m+1}(\mathbf{X}) - \boldsymbol{\theta}_{m+1}^*(\mathbf{X}))^2 = \boldsymbol{\theta}_{m+1}^*(\mathbf{X})^2 \leq \tilde{c}_6 \cdot \|g(\boldsymbol{\theta}^*, \mathbf{X})\|^2 = \tilde{c}_6 \cdot \|g(\boldsymbol{\theta}, \mathbf{X}) - g(\boldsymbol{\theta}^*, \mathbf{X})\|^2.$$

If $\mathbf{X} \notin \mathcal{E} \cup \mathcal{E}_1$, it must be that there is $\ell \in \{1, 2, 3, 4\}$, such that $|g_\ell(\boldsymbol{\theta}, \mathbf{X}) - g_\ell(\boldsymbol{\theta}^*, \mathbf{X})| > \epsilon$. This clearly implies that for some $\tilde{c}_7 > 0$,

$$(\boldsymbol{\theta}_{m+1}(\mathbf{X}) - \boldsymbol{\theta}_{m+1}^*(\mathbf{X}))^2 \leq \tilde{c}_7 \cdot \|g(\boldsymbol{\theta}, \mathbf{X}) - g(\boldsymbol{\theta}^*, \mathbf{X})\|^2.$$

by the boundedness of $\boldsymbol{\theta}_{m+1}(\mathbf{X})$ and $\boldsymbol{\theta}_{m+1}^*(\mathbf{X})$. Combined, we have that

$$\mathbb{E}_{\mathbf{X}} \left[(\boldsymbol{\theta}_{m+1}(\mathbf{X}) - \boldsymbol{\theta}_{m+1}^*(\mathbf{X}))^2 \right] \leq \max\{\tilde{c}_4, \tilde{c}_6, \tilde{c}_7\} \cdot \mathbb{E}_{\mathbf{X}} \left[\|g(\boldsymbol{\theta}, \mathbf{X}) - g(\boldsymbol{\theta}^*, \mathbf{X})\|^2 \right].$$

Note that (27) implies there exists a constant $\tilde{c}_7 > 0$ such that,

$$\begin{aligned} \mathbb{E}_{\mathbf{X}, \mathbf{T}} \left[(G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{T}) - G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{T}))^2 \right] &\geq \tilde{c}_8 \mathbb{E}_{\mathbf{X}} \left[(G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_1 \cup \{i\})) - G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t}(S_1 \cup \{i\})))^2 \right] \\ \mathbb{E}_{\mathbf{X}, \mathbf{T}} \left[(G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{T}) - G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{T}))^2 \right] &\geq \tilde{c}_8 \mathbb{E}_{\mathbf{X}} \left[(G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_1)) - G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t}(S_1)))^2 \right] \\ \mathbb{E}_{\mathbf{X}, \mathbf{T}} \left[(G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{T}) - G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{T}))^2 \right] &\geq \tilde{c}_8 \mathbb{E}_{\mathbf{X}} \left[(G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_2 \cup \{i\})) - G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t}(S_2 \cup \{i\})))^2 \right] \\ \mathbb{E}_{\mathbf{X}, \mathbf{T}} \left[(G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{T}) - G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{T}))^2 \right] &\geq \tilde{c}_8 \mathbb{E}_{\mathbf{X}} \left[(G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{t}(S_2)) - G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t}(S_2)))^2 \right]. \end{aligned} \tag{31}$$

Then,

$$\mathbb{E}_{\mathbf{X}, \mathbf{T}} \left[(G(\boldsymbol{\theta}(\mathbf{X}), \mathbf{T}) - G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{T}))^2 \right] \geq \frac{\tilde{c}_8}{4} \cdot \frac{1}{\max\{\tilde{c}_4, \tilde{c}_6, \tilde{c}_7\}} \cdot \mathbb{E}_{\mathbf{X}} \left[(\boldsymbol{\theta}_{m+1}(\mathbf{X}) - \boldsymbol{\theta}_{m+1}^*(\mathbf{X}))^2 \right]. \tag{32}$$

The rest of the proof is the same as that for the *Generalized Sigmoid Form I*. \square

A.5. Proof of Proposition 2

Proof. The results directly follow from Theorem 2 of Farrell et al. (2020), which we also restate as Lemma 1 in Appendix A.3.

First, we derive the first-order derivative $\ell_{\boldsymbol{\theta}}$ and the Hessian matrix $\ell_{\boldsymbol{\theta}\boldsymbol{\theta}}$. With the loss function $\ell(y, \mathbf{t}, \boldsymbol{\theta}(\mathbf{x})) = (y - G(\boldsymbol{\theta}(\mathbf{x}), \mathbf{t}))^2$, we can compute to get

$$\ell_{\boldsymbol{\theta}} = (\partial \ell / \partial \theta_1, \dots, \partial \ell / \partial \theta_{d_{\boldsymbol{\theta}}})' = 2(G - y)G_{\boldsymbol{\theta}},$$

$$\ell_{\boldsymbol{\theta}\boldsymbol{\theta}} = 2G_{\boldsymbol{\theta}}G'_{\boldsymbol{\theta}} + 2G_{\boldsymbol{\theta}\boldsymbol{\theta}}(G - y),$$

$$\Lambda(\mathbf{x}) = \mathbb{E}[\ell_{\boldsymbol{\theta}\boldsymbol{\theta}} | \mathbf{X} = \mathbf{x}] = 2\mathbb{E}_{\mathbf{T}|\mathbf{x}}[G_{\boldsymbol{\theta}}G'_{\boldsymbol{\theta}} | \mathbf{x}] + 2\mathbb{E}_{\mathbf{T}|\mathbf{x}} \left[G_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbb{E}_{y|\mathbf{t}, \mathbf{x}}[(G - y) | \mathbf{T} = \mathbf{t}, \mathbf{X} = \mathbf{x}] | \mathbf{X} = \mathbf{x} \right] = 2\mathbb{E}[G_{\boldsymbol{\theta}}G'_{\boldsymbol{\theta}} | \mathbf{X} = \mathbf{x}].$$

We now verify the assumptions of Lemma 1. The first condition, identification of $\boldsymbol{\theta}^*$, is satisfied under Assumptions 1 and 4, which is shown in Proposition 1(a). The second condition holds by Assumption 2(i) (i.e., DGP (1) holds) and the mean squared loss function. The third condition follows from Assumption 2(ii). The fourth condition regarding the pathwise differentiable follows from Assumption 2(iii), and the thrice continuously differentiability holds by the form of G defined in Assumption 1 and the mean squared loss with sufficient smoothness in $\boldsymbol{\theta}$. The fourth condition holds because the link functions G in Assumption 1 are sufficiently smooth together with the boundedness of $\boldsymbol{\theta}^*(\mathbf{X})$ given by Assumption 3. Thus, Proposition 2 follows immediately from Lemma 1. \square

A.6. Verification of Assumption 2(ii)

In the following, we illustrate that, if the treatment assignment mechanism $\nu(\cdot | \cdot)$ satisfies Assumption 4, Assumption 2(ii) (i.e., $\Lambda(\mathbf{x}) \succ 0$) is easy to satisfy and can be translated into very lenient conditions. $\Lambda(\mathbf{x}) \succ 0$ is equivalent to that the vectors $\{G_{\theta}(\boldsymbol{\theta}(\mathbf{x}), \mathbf{t})\}_{\mathbf{t}}$ are nondegenerate. Without loss of generality, we take the *Generalized Sigmoid Form II* for illustration. The other forms of link function can be verified similarly.

By definition, we have $\Lambda(\mathbf{x}) = 2\mathbb{E}[G_{\theta}(\boldsymbol{\theta}(\mathbf{x}), \mathbf{T})G_{\theta}(\boldsymbol{\theta}(\mathbf{x}), \mathbf{T})' | \mathbf{X} = \mathbf{x}]$, where

$$G_{\theta}(\boldsymbol{\theta}(\mathbf{x}), \mathbf{t})' = \left(\frac{\theta_{m+1}(\mathbf{x}) \exp(-(\theta_0(\mathbf{x}) + \dots + \theta_m(\mathbf{x})t_m))}{(1 + \exp(-(\theta_0(\mathbf{x}) + \dots + \theta_m(\mathbf{x})t_m)))^2} (1, t_1, \dots, t_m), \frac{1}{1 + \exp(-(\theta_0(\mathbf{x}) + \dots + \theta_m(\mathbf{x})t_m))} \right).$$

Thus, to verify that $\Lambda(\mathbf{x})$ is invertible, it suffices to show that the matrix constructed by vectors $\{G_{\theta}(\boldsymbol{\theta}(\mathbf{x}), \mathbf{t})\}_{\mathbf{t}}$ has full rank $m + 2$. For ease of exposition, we drop the dependence on \mathbf{x} , which will not cause any confusion.

Next, we verify that $\Lambda(\mathbf{x}) \succ 0$ for the treatment assignment mechanism in Assumption 4(b-iv). All other assignment rules in Assumption 4(b) can be translated into similar lenient conditions, so we omit their verification for brevity. Specifically, to ensure that $\mathbb{E}[\tilde{\mathbf{T}}\tilde{\mathbf{T}}' | \mathbf{X} = \mathbf{x}] \succ 0$, we consider the assignment mechanism $\nu(\mathbf{t}(\emptyset) | \mathbf{x}) > 0$, $\nu(\mathbf{t}(\{1\}) | \mathbf{x}) > 0$, $\nu(\mathbf{t}(\{2\}) | \mathbf{x}) > 0, \dots$, $\nu(\mathbf{t}(\{m\}) | \mathbf{x}) > 0$, together with the overlapping assignment $\nu(\mathbf{t}(S_1) | \mathbf{x}) > 0$, $\nu(\mathbf{t}(S_1 \cup \{i\}) | \mathbf{x}) > 0$, $\nu(\mathbf{t}(S_2) | \mathbf{x}) > 0$, and $\nu(\mathbf{t}(S_2 \cup \{i\}) | \mathbf{x}) > 0$. Define $e_0 := \exp(-\theta_0)$, $e_i := \exp(-\theta_0 - \theta_i)$, for $i = 1, 2, \dots, m$, and $e_{m+1} := \exp(-\boldsymbol{\theta}'\mathbf{t}(S_1 \cup \{i\}))$. We have

$$\begin{aligned} \text{rank}(\Lambda) &\stackrel{(a)}{\geq} \text{rank} \left(\begin{array}{cc} \frac{\theta_{m+1}e_0}{1+e_0}(1, 0, 0, 0, \dots, 0) & \frac{1}{1+e_0} \\ \frac{\theta_{m+1}e_1}{1+e_1}(1, 1, 0, 0, \dots, 0) & \frac{1}{1+e_1} \\ \frac{\theta_{m+1}e_2}{1+e_2}(1, 0, 1, 0, \dots, 0) & \frac{1}{1+e_2} \\ \frac{\theta_{m+1}e_3}{1+e_3}(1, 0, 0, 1, \dots, 0) & \frac{1}{1+e_3} \\ \vdots & \vdots \\ \frac{\theta_{m+1}e_m}{1+e_m}(1, 0, 0, 0, \dots, 1) & \frac{1}{1+e_m} \\ \frac{\theta_{m+1}e_{m+1}}{1+e_{m+1}}(1, \mathbf{t}(S_1 \cup \{i\})') & \frac{1}{1+e_{m+1}} \end{array} \right) \stackrel{(b)}{=} \text{rank} \left(\begin{array}{cc} \frac{e_0}{1+e_0}(1, 0, 0, 0, \dots, 0) & \frac{1}{1+e_0} \\ \frac{e_1}{1+e_1}(1, 1, 0, 0, \dots, 0) & \frac{1}{1+e_1} \\ \frac{e_2}{1+e_2}(1, 0, 1, 0, \dots, 0) & \frac{1}{1+e_2} \\ \frac{e_3}{1+e_3}(1, 0, 0, 1, \dots, 0) & \frac{1}{1+e_3} \\ \vdots & \vdots \\ \frac{e_m}{1+e_m}(1, 0, 0, 0, \dots, 1) & \frac{1}{1+e_m} \\ \frac{e_{m+1}}{1+e_{m+1}}(1, \mathbf{t}(S_1 \cup \{i\})') & \frac{1}{1+e_{m+1}} \end{array} \right) \\ &\stackrel{(c)}{=} \text{rank} \left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & & & \vdots \\ 1 & 0 & 0 & 1 \\ 1 & & & \mathbf{t}(S_1 \cup \{i\})' \end{array} \begin{array}{c} \dots \\ 0 \\ \dots \\ 0 \\ \dots \\ 1 \\ \frac{1}{e_{m+1}} \end{array} \right) \stackrel{(d)}{=} \text{rank} \left(\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array} \begin{array}{c} \dots \\ 0 \\ \dots \\ 0 \\ \dots \\ 1 \\ \frac{1}{e_{m+1}} - \frac{1}{e_0} - (\frac{1}{e_1} - \frac{1}{e_0}) - (\frac{1}{e_2} - \frac{1}{e_0}) - \dots - (\frac{1}{e_i} - \frac{1}{e_0}) \end{array} \right). \end{aligned}$$

The inequality (a) follows from the fact the right-hand-side matrix is expanded only by a partial of vectors $\{G_{\theta}(\boldsymbol{\theta}(\mathbf{x}), \mathbf{t})\}_{\mathbf{t}}$. The equality (b) follows from $\theta_{m+1} \neq 0$. The equality (c) follows from $1 + e_i \neq 0$ and $e_i \neq 0$ for all $i = 1, 2, \dots, m$. The equality (d) follows from some subtraction operations by rows. To guarantee the full rank condition that $\text{rank}(\Lambda) = m + 2$, a sufficient condition is $\frac{1}{e_{m+1}} - \frac{1}{e_1} - \frac{1}{e_i} + \frac{1}{e_0} \neq 0$, i.e., the bottom-right entry of the last matrix is nonzero. This is indeed a very weak condition.

For other assignment mechanisms in Assumption 4(b), one can also translate the invertibility of Λ into very lenient conditions. We omit the details for brevity.

A.7. ATE Estimator Construction via Cross-Fitting

In the following, we introduce the sample-splitting/cross-fitting and estimation procedure from Chernozhukov et al. (2018). The entire DeDL framework is given by Algorithm 1. In our setting with the known distribution of \mathbf{T} , we require only the two-way splitting, but for unknown \mathbf{T} distribution, three-way splitting is required, with an additional portion of samples used for obtaining $\hat{\Lambda}(\mathbf{x})$.

One can split the data samples $\{1, 2, \dots, n\}$ into S nonoverlapping copies $\mathcal{S}_s \subset \{1, 2, \dots, n\}$, $s = 1, 2, \dots, S$ with the cardinality $|\mathcal{S}_s|$ being proportionally to the sample size n , and let \mathcal{S}_s^c be the complement of \mathcal{S}_s . First, we use \mathcal{S}^c to get estimators $\hat{\theta}_s(\cdot)$ of parameters $\theta^*(\cdot)$, and compute $\hat{\Lambda}_s(\cdot)$ given the estimators $\hat{\theta}_s(\cdot)$ and distribution of \mathbf{T} . Then, one can use the other samples to construct an estimator of $\mu(\mathbf{t})$, for any $\mathbf{t} \in \{0, 1\}^m$ as

$$\hat{\mu}_{\text{DeDL}}(\mathbf{t}) = \frac{1}{S} \hat{\mu}_s(\mathbf{t}), \quad \hat{\mu}_s(\mathbf{t}) = \frac{1}{|\mathcal{S}_s|} \sum_{i \in \mathcal{S}_s} \psi(\mathbf{z}_i, \hat{\theta}_s(\mathbf{x}_i), \hat{\Lambda}_s(\mathbf{x}_i); \mathbf{t}, \mathbf{t}_0). \quad (33)$$

Similarly, the variance estimator can be constructed as

$$\hat{\Psi}_{\text{DeDL}}(\mathbf{t}; \mu) = \frac{1}{S} \hat{\Psi}_s(\mathbf{t}), \quad \hat{\Psi}_s(\mathbf{t}) = \frac{1}{|\mathcal{S}_s|} \sum_{i \in \mathcal{S}_s} \left(\psi(\mathbf{z}_i, \hat{\theta}_s(\mathbf{x}_i), \hat{\Lambda}_s(\mathbf{x}_i); \mathbf{t}, \mathbf{t}_0) - \hat{\mu}_{\text{DeDL}}(\mathbf{t}) \right)^2. \quad (34)$$

The asymptotic normality of $\hat{\mu}_{\text{DeDL}}(\mathbf{t})$ in Theorem 2(a) directly follows from Chernozhukov et al. (2018), and a detailed proof can also be found in Farrell et al. (2020).

Therefore, the $(1 - \alpha)$ -confidence interval of $\hat{\mu}_{\text{DeDL}}(\mathbf{t})$ is given by

$$\widehat{\mathcal{C}I}_{\text{DeDL}}(\mathbf{t}; \mu) = \left[\hat{\mu}_{\text{DeDL}}(\mathbf{t}) - \frac{1}{\sqrt{n}} \cdot \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \cdot \hat{\Psi}_{\text{DeDL}}(\mathbf{t}; \mu), \hat{\mu}_{\text{DeDL}}(\mathbf{t}) + \frac{1}{\sqrt{n}} \cdot \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \cdot \hat{\Psi}_{\text{DeDL}}(\mathbf{t}; \mu) \right], \quad (35)$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of a standard normal random variable.

A.8. Estimation and Inference for Best Treatment Identification

Details for the best treatment identification. After obtaining the asymptotically normal estimators of ATE $\hat{\mu}(\mathbf{t})$ for all experiment combinations $\mathbf{t} \in \{0, 1\}^m$, the next step is identifying the best experiment combination, which is defined as $\mathbf{t}^* := \arg \max_{\mathbf{t} \in \{0, 1\}^m} \mu(\mathbf{t})$.

Following the common practice, one can search for the best treatment combination by searching for the highest ATE estimation. Formally, we define the empirical best treatment level as $\hat{\mathbf{t}}^* := \arg \max_{\mathbf{t} \in \{0, 1\}^m} \hat{\mu}_{\text{DeDL}}(\mathbf{t})$. The remaining job is verifying whether $\hat{\mathbf{t}}^*$ is the best treatment level with significant improvements over all other treatment levels, i.e., we test the one-sided hypothesis on whether the ATE of $\hat{\mathbf{t}}^*$ is significantly better than other treatments:

$$H_0: \tau(\mathbf{t}) > 0, \text{ for all } \mathbf{t} \in \{0, 1\}^m \setminus \{\hat{\mathbf{t}}^*\},$$

where

$$\tau(\mathbf{t}) := \mu(\hat{\mathbf{t}}^*) - \mu(\mathbf{t}) = \mathbb{E}[G(\boldsymbol{\theta}^*(\mathbf{X}), \hat{\mathbf{t}}^*)] - \mathbb{E}[G(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{t})],$$

is the improvement in ATE of the empirically optimal $\hat{\mathbf{t}}^*$ over treatment level $\mathbf{t} \in \{0, 1\}^m$. Notice that the $\tau(\mathbf{t})$ can be rewritten as,

$$\tau(\mathbf{t}) = \mathbb{E}[H(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}); \hat{\mathbf{t}}^*, \mathbf{t})],$$

which is similar in structure to $\mu(\mathbf{t}) = \mathbb{E}[H(\mathbf{X}, \boldsymbol{\theta}^*(\mathbf{X}); \mathbf{t}, \mathbf{t}_0)]$ with a change of the inputs in the advantage function H .

Applying an influence function similar to (10) developed in Section 3.4, i.e., $\psi(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\Lambda}; \hat{\mathbf{t}}^*, \mathbf{t}) = H(\mathbf{x}, \boldsymbol{\theta}(\mathbf{x}); \hat{\mathbf{t}}^*, \mathbf{t}) - H_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\theta}(\mathbf{x}); \hat{\mathbf{t}}^*, \mathbf{t})' \boldsymbol{\Lambda}(\mathbf{x})^{-1} \ell_{\boldsymbol{\theta}}(y, \check{\mathbf{t}}, \boldsymbol{\theta}(\mathbf{x}))$ and the two-way splitting procedure described above, we construct the estimators

$$\hat{\tau}_{\text{DeDL}}(\mathbf{t}) := \hat{\mu}_{\text{DeDL}}(\hat{\mathbf{t}}^*_{\text{DeDL}}) - \hat{\mu}_{\text{DeDL}}(\mathbf{t}), \quad (36)$$

and the variance estimate $\hat{\Psi}_{\text{DeDL}}(\mathbf{t}; \tau)$ for $\tau_{\text{DeDL}}(\mathbf{t})$.

Proof of Theorem 2

By Proposition 1(b), it is straightforward to verify that $\|\hat{\theta}_{sk} - \theta_k^*\|_{L_2(\mathbf{X})} = o(n^{-1/4})$ holds if $p > d_{\mathbf{X}}$. By Theorem 3 of Farrell et al. (2020), we have the desired asymptotic normality of the proposed estimators $\hat{\mu}_{\text{DeDL}}(\mathbf{t})$ and $\hat{\tau}_{\text{DeDL}}(\mathbf{t})$. This proves part (a). On the other hand, we also highlight that this DNN convergence rate is inherited from Lemma 1 (Theorem 2 in Farrell et al. 2020), which might not be tight as well. As establishing a tighter rate for DNN convergence is beyond the scope of our paper, we instead follow the standard $o(n^{-1/4})$ decay rate (as claimed in Proposition 1).

For Part (b), the empirically optimal treatment combination $\hat{\mathbf{t}}^* := \arg \max_{\mathbf{t} \in \{0,1\}^m} \hat{\mu}(\mathbf{t})$ depends on the samples used for the training and inference of ATE, which is a critical challenge in our proof below. For the rest of our proof, we drop the subscript DeDL to simplify the notation. First, we construct the following estimator $\hat{\zeta}(\mathbf{t}) := \hat{\mu}(\mathbf{t}^*) - \hat{\mu}(\mathbf{t})$, where $\hat{\mu}(\cdot)$ is given by (33), for the true advantage of the optimal treatment combination $\zeta(\mathbf{t}) := \mu(\mathbf{t}^*) - \mu(\mathbf{t})$. Similar to (34), one can construct the variance estimator $\hat{\Psi}(\mathbf{t}; \zeta)$ for $\hat{\zeta}(\mathbf{t})$. Importantly, $\hat{\zeta}(\mathbf{t})$ is a virtual estimator not available in practice, because the ground-truth optimal treatment combination \mathbf{t}^* is unknown.

First, Theorem 3 in Farrell et al. (2020) establishes the asymptotic normality of $\hat{\zeta}(\mathbf{t})$:

$$\sqrt{n}(\hat{\Psi}(\mathbf{t}; \zeta))^{-1/2}(\hat{\zeta}(\mathbf{t}) - \zeta(\mathbf{t})) = \sum_{i=1}^n (\hat{\Psi}(\mathbf{t}; \zeta))^{-1/2}(\psi(\mathbf{z}_i, \boldsymbol{\theta}^*(\mathbf{x}_i), \boldsymbol{\Lambda}(\mathbf{x}_i), \mathbf{t}^*, \mathbf{t}) - \zeta(\mathbf{t})) / \sqrt{n} + o_p(1) \rightarrow_d \mathcal{N}(0, 1). \quad (37)$$

Next, we show that, with probability going to 1, the optimal treatment combination is correctly identified, i.e., $\lim_{n \rightarrow \infty} \mathbb{P}[\hat{\mathbf{t}}^* = \mathbf{t}^*] = 1$, when the dependence on sample size n is dropped for brevity. By Part (a), we have $\hat{\mu}(\mathbf{t}^*) \rightarrow_p \mu(\mathbf{t}^*)$ and $\hat{\mu}(\mathbf{t}) \rightarrow_p \mu(\mathbf{t})$ for any $\mathbf{t} \neq \mathbf{t}^*$. Because the max operator is continuous with respect to the L_1 norm, we have

$$|\hat{\mu}(\mathbf{t}^*) - \mu(\mathbf{t}^*)| \vee |\hat{\mu}(\mathbf{t}) - \mu(\mathbf{t})| \rightarrow_p 0,$$

where “ \vee ” denote the maximum of two real numbers. Since the optimal treatment combination \mathbf{t}^* is unique, as the sample size n goes to infinity,

$$|\hat{\mu}(\mathbf{t}^*) - \mu(\mathbf{t}^*)| \vee |\hat{\mu}(\mathbf{t}) - \mu(\mathbf{t})| \leq \frac{\mu(\mathbf{t}^*) - \mu(\mathbf{t})}{4}, \quad (38)$$

with a probability going to 1.

Furthermore, (38) implies,

$$\hat{\mu}(\mathbf{t}^*) - \hat{\mu}(\mathbf{t}) = [\mu(\mathbf{t}^*) - (\mu(\mathbf{t}^*) - \hat{\mu}(\mathbf{t}^*))] - [\mu(\mathbf{t}) - (\mu(\mathbf{t}) - \hat{\mu}(\mathbf{t}))] \geq \frac{\mu(\mathbf{t}^*) - \mu(\mathbf{t})}{2} > 0.$$

Therefore, with the probability going to 1, we have $\hat{\mu}(\mathbf{t}^*) > \hat{\mu}(\mathbf{t})$ as the sample size n goes to infinity. Taking the union bound over $(2^m - 1)$ treatment combinations, we have $\hat{\mu}(\mathbf{t}^*) > \max_{\mathbf{t} \neq \mathbf{t}^*} \hat{\mu}(\mathbf{t})$, i.e., $\hat{\mathbf{t}}^* = \arg \max_{\mathbf{t}} \hat{\mu}(\mathbf{t}) = \mathbf{t}^*$, with a probability going to 1 as the sample size n goes to infinity.

If $\hat{\mathbf{t}}^* = \mathbf{t}^*$, by definition,

$$\sqrt{n}(\hat{\Psi}(\mathbf{t}; \zeta))^{-1/2}(\hat{\zeta}(\mathbf{t}) - \zeta(\mathbf{t})) - \sqrt{n}(\hat{\Psi}(\mathbf{t}; \tau))^{-1/2}(\hat{\tau}(\mathbf{t}) - \tau(\mathbf{t})) = 0$$

Hence, we have that for any $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}[|\sqrt{n}(\hat{\Psi}(\mathbf{t}; \zeta))^{-1/2}(\hat{\zeta}(\mathbf{t}) - \zeta(\mathbf{t})) - \sqrt{n}(\hat{\Psi}(\mathbf{t}; \tau))^{-1/2}(\hat{\tau}(\mathbf{t}) - \tau(\mathbf{t}))| < \varepsilon] = 1. \quad (39)$$

Combining (39) with the asymptotic normality (37), by Slutsky's theorem, we have

$$\sqrt{n}(\hat{\Psi}(\mathbf{t}; \tau))^{-1/2}(\hat{\tau}(\mathbf{t}) - \tau(\mathbf{t})) = \sum_{i=1}^n (\hat{\Psi}(\mathbf{t}; \tau))^{-1/2}(\psi(\mathbf{z}_i, \boldsymbol{\theta}^*(\mathbf{x}_i), \boldsymbol{\Lambda}(\mathbf{x}_i), \mathbf{t}^*, \mathbf{t}) - \tau(\mathbf{t})) / \sqrt{n} + o_p(1) \xrightarrow{d} \mathcal{N}(0, 1),$$

which concludes the proof. \square

Given the asymptotic normality, the $(1 - \alpha)$ -confidence interval for $\hat{\tau}_{\text{DeDL}}(\mathbf{t})$ is given by

$$\widehat{\mathcal{CI}}_{\text{DeDL}}(\mathbf{t}; \tau) = \left[\hat{\tau}_{\text{DeDL}}(\mathbf{t}) - \frac{1}{\sqrt{n}} \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \hat{\Psi}_{\text{DeDL}}(\mathbf{t}; \tau), \hat{\tau}_{\text{DeDL}}(\mathbf{t}) + \frac{1}{\sqrt{n}} \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \hat{\Psi}_{\text{DeDL}}(\mathbf{t}; \tau) \right]. \quad (40)$$

Appendix B: Empirical Analysis With Field Experiment Data

B.1. User Covariates Used in Section 4

Table A1 presents all the user covariates used in our empirical analysis in Section 4.

B.2. Stratified Sampling Details

We employ a three-step stratified sampling to keep the user covariates balanced with respect to eight different treatment combinations. First, we categorize 10 continuous variables in Table A1 by their quantile ranges [0%, 25%), [25%, 50%), [50%, 75%), and [75%, 100%]. Specifically, we assign 1, 2, 3, and 4 as new values for values in each quantile interval respectively for each continuous variable. After all the variables are discretized, we proceed to divide the population into subpopulations according to imbalanced covariates. To check for diversely distributed covariates, we utilize a pairwise t-test between the baseline combination (0, 0, 0) and seven treatment combinations for covariates in Table A1. Among the 26 covariates, 16 discrete covariates show no significant difference between the baseline combination (0, 0, 0) and the other seven treatment combinations, while 10 continuous variables are imbalanced-distributed. Therefore, we divide users into 69,111 strata by the value of imbalanced covariates. Namely, users with the same value in all imbalanced covariates are grouped together. For each stratum, we set the minimum number of users among all treatment combinations as its target size and then randomly sample this many individuals for each combination as the stratified sample. Therefore, the stratified sample has a similar number of users in each treatment combination, and the treatment assignment mechanism satisfies $\mathbb{P}[T_i = 1 | \mathbf{X} = \mathbf{x}] = \mathbb{P}[T_i = 0 | \mathbf{X} = \mathbf{x}] = 0.5$ ($i = 1, 2, 3$) for any \mathbf{x} .

B.3. Complete Randomization Check Results

Table A2 presents complete randomization check results of the stratified samples used in our empirical analysis in Section 4.

Table A1 User Covariates Used in the DeDL Framework

	Variable	Description
Disc- rete Var.	Age Range	Age range of user: young, mid-age, senior, old, or unknown
	Gender	Gender of user: male, female, or unknown
	Operating System	OS of user's device: Android, IOS, or other
	Product Version	Version used: lite, express, or regular
	Phone Price Range	Price range of device: luxury, expensive, affordable, or unknown
	User Activeness Degree	Activeness of user: high-, mid-, low-active, or new user
	User Active-Deepness Degree	Active-deepness of user: deep-, shallow-active, or new user
	Feed Mode	Preferred mode of app interface: video stream, video cover stream, or unknown
	Number of Followers	Interval of the user's number of followers: <10, 10 - 10k, 10k - 100k, >100k
	Influencer Level	Mainly determined by the number of followers: micro, midtier, or macro influencer
Conti- nuous Var.	Number of Mutual Followers	Interval of the user's number of mutual followers (friends): <10, 10 - 10k, 10k - 100k, >100k
	Sociableness Level	Mainly determined by the number of mutual followers (friends): low-, mid-, or high-sociable
	Frequent Residence Area	Region in which the user is frequently on the platform: South, North, or unknown
	Frequent Residence City Level	Level of city in which the user is frequently on the platform: large city, big city, medium city, small city, or unknown
	Frequent Residence City Type	Type of city in which the user is frequently on the platform: city, town, rural, or unknown
	Active Engagement City Level	Level of city in which the user is always active: large city, big city, medium city, small city, or unknown
	Average App Usage Duration	User's average usage duration on platform per day
	Average Video Watching Time	User's average time on watching videos on platform per day
User De- mographics	Average Live Watching Time	User's average time on watching live on platform per day
	Average DP Video Watching Time	User's average time on watching videos on Discover Page per day
	Average LP Video Watching Time	User's average time on watching videos on Live Page per day
	Average FYP Video Watching Time	User's average time on watching videos on For You Page per day
	Average FP Video Watching Time	User's average time on watching videos on Following Page per day
	Average DP Screen Time	User's average time on Discover Page per day
	Average LP Screen Time	User's average time on Live Page per day
	Average FP Screen Time	User's average time on Following Page per day

Table A2 Randomization Check

	Treatment Combination t'	(0,0,0)	(0,0,1)	(0,1,0)	(1,0,0)	(1,1,1)	(1,1,0)	(1,0,1)	(0,1,1)
User De- mographics	Proportion of Male Users	60.51% (0.41)	60.63% (0.13)	60.31% (0.85)	60.49% (0.40)	60.40% (0.26)	60.36% (0.11)	60.30% (0.15)	60.31% (0.15)
	Proportion of High-Active Users	29.67% (0.88)	29.65% (0.55)	29.60% (0.62)	29.73% (0.44)	29.77% (0.85)	29.65% (0.34)	29.80% (0.91)	29.66% (0.91)
	Proportion of Users from the South	39.83% (0.16)	39.64% (0.63)	39.90% (0.88)	39.80% (0.90)	39.85% (0.26)	39.68% (0.29)	39.69% (0.48)	39.74% (0.48)
	Average Active Days per User	-0.0003 (0.78)	-0.0010 (0.59)	0.0012 (0.91)	-0.0006 (0.75)	-0.0011 (0.88)	0.0001 (0.61)	0.0011 (0.69)	0.0008 (0.69)
	Average DP Screen Time per User	0.0173 (0.31)	0.0144 (0.45)	0.0151 (0.46)	0.0152 (0.25)	0.0140 (0.34)	0.0145 (0.46)	0.0152 (0.58)	0.0157 (0.58)
User Behaviors 10 Days Prior to the Experi- ments	Average LP Screen Time per User	0.0126 (0.21)	0.0091 (0.49)	0.0146 (0.42)	0.0104 (0.75)	0.0136 (0.52)	0.0145 (0.42)	0.0103 (0.25)	0.0159 (0.25)
	Average FYP Screen Time per User	0.0048 (0.36)	0.0022 (0.62)	0.0033 (0.91)	0.0044 (0.18)	0.0010 (0.61)	0.0033 (0.25)	0.0015 (0.41)	0.0024 (0.41)
	Average App Usage Duration per User	0.0129 (0.55)	0.0111 (0.62)	0.0114 (0.61)	0.0114 (0.39)	0.0104 (0.37)	0.0103 (0.76)	0.0120 (0.69)	0.0117 (0.69)

Note: p-values of t-tests between the baseline combination $t_0 = (0,0,0)'$ and other treatment combinations are reported in parentheses. To protect sensitive data, the reported metrics of active days, screen time, and app-usage duration are rescaled.

Appendix C: Implementation Details of Benchmarks

In this section, we present the implementation details of using linear addition (LA), linear regression (LR), pure deep learning (PDL), and structured deep learning (SDL) as our benchmark to estimate ATE. For LR, PDL, and SDL estimation, we follow the same four-fold cross-fitting described in Appendix 5.2.

C.1. Linear Addition (LA) Estimator

To obtain the LA estimator for each treatment level, we use the ATE of five observed treatment combinations to calculate the ATE of unobserved treatment combinations. For observed treatment combinations, their estimated ATE and significance level are the same as the ground truth. For unobserved treatment combinations, we use the linear addition of ATE of individual observed treatment combinations as the estimated ATE. Furthermore, we estimate the standard error of the estimated ATE for unobserved treatment combinations by assuming that the estimators for individual experiments are independent.

We report the LA estimators in the first row in each section of treatment combination in Table A3. The top four treatment combinations are observable, and therefore the LA estimator yields zero error (see columns 5–7 in Table A3). The estimators for the ATE of the bottom-three treatment combinations use the ATE of individual treatment combinations, i.e., $(0, 0, 1)$, $(0, 1, 0)$, and $(1, 0, 0)$, to calculate the final results $\hat{\mu}(\mathbf{t}) = \hat{\mu}(\mathbf{t}_1) + \hat{\mu}(\mathbf{t}_2)$, where $\mathbf{t} = \mathbf{t}_1 + \mathbf{t}_2$. The standard deviation of the estimator follows $\hat{\sigma}(\hat{\mu}(\mathbf{t})) = \sqrt{\hat{\sigma}(\hat{\mu}(\mathbf{t}_1))^2 + \hat{\sigma}(\hat{\mu}(\mathbf{t}_2))^2}$.

C.2. Linear Regression (LR) Estimator

The LR estimator uses the regression coefficients as the estimated ATE. The regression is defined as

$$y = \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3 + \alpha \mathbf{x}, \quad (41)$$

where t_1 , t_2 , and t_3 denote three experiments for pages DP, LP, and FYP, respectively, and \mathbf{x} denotes the user covariates in Table A1. We fit the linear regression model (41) with 1,291,652 data points with the five observed treatment combinations. The estimator for β_i , $\hat{\beta}_i$, captures the ATE of treatment t_i . Similar to LA estimators, LR estimators assume the linear additivity of the ATEs. To maintain a fair comparison between the estimators, we adopt a four-fold cross-fitting described in Appendix 5.2 to obtain the LR estimator as well. Specifically, we fit the linear regression model with 75% of the training data. For the second-stage inference, we predict the potential outcome of each user given the covariates \mathbf{x} under treatment combination $\mathbf{t} \in \mathcal{T}$ using the trained linear regression. The estimation and inference of each ATE is obtained through the standard pairwise t-test for each treatment combination with the predicted outcomes on the last data fold. The average values of estimated ATEs for each treatment combination are presented in the second row in Table A3.

C.3. Pure Deep Learning (PDL) Estimator

Similar to the implementation of DeDL approach in Appendix 5.2, we apply the four-fold cross-fitting procedure to reduce overfitting when implementing the PDL estimators. Specifically, we evenly partition the observed data of five treatment combinations into four random folds. For each data fold, we use the data from three other folds as the training data to approximate $y = f(\mathbf{x}, \mathbf{t})$, where $f(\cdot, \cdot)$ is a fully-connected

three-layer DNN with 20 nodes in each layer. We call this approach the “pure” deep learning estimator because no constraints are imposed on the DNN $f(\cdot, \cdot)$. The DNN is trained with the Adam optimizer and the mean squared error as the loss function. We then use the trained DNN $\hat{f}(\cdot, \cdot)$ to predict the potential outcomes under eight treatment combinations for each user the covariates \mathbf{x} on the last fold. Similar to the LR estimator, pairwise t-tests with the predicted outcomes between any treatment combination \mathbf{t} and the control \mathbf{t}_0 provide the ATE estimate and its corresponding standard error, whereas the final estimates are the average value of all four data folds. The fourth row for each treatment combination in Table A3 shows the estimated ATE of the PDL approach.

C.4. Structured Deep Learning (SDL) Estimator

The SDL approach adopts the same structured DNN and four-fold cross-fitting as DeDL as documented in Appendix 5.2, but without the debiasing term. For each treatment combination $\mathbf{t} \in \mathcal{T}$, we use the trained structured DNN to predict the potential outcome given the covariates \mathbf{x} : $H(\mathbf{x}, \hat{\theta}_s(\mathbf{x}); \mathbf{t}, \mathbf{t}_0) := G(\hat{\theta}_s(\mathbf{x}), \mathbf{t}) - G(\hat{\theta}_s(\mathbf{x}), \mathbf{t}_0)$. Similar to the LA and PDL approaches, t-tests with the predicted outcomes provide the ATE estimates under SDL. We remark that the SDL approach underestimates the standard errors because it ignores the noises in fitting $\hat{\theta}(\cdot)$ caused by the variations in the training data. The estimation results of SDL for each treatment combination are reported in the last row in Table A3.

Appendix D: Details of Synthetic Experiments

In this section, we provide details of the synthetic experiments designed to assess several factors we have identified that could potentially influence the performance of the DeDL estimators in practice. We first validate our theory by varying the number of experiments $m \in \{4, 6, 8, 10\}$ in Appendix D.2. In Appendix D.3, we test the performance of our DeDL estimators with a potentially large bias of estimating $\hat{\theta}(\cdot)$; we find that DeDL is fairly robust, with moderate biases. We also systematically assess the performance of DeDL under model misspecification, and we shed light on how to test and select the link function in practice in Appendix D.4. Furthermore, in Appendix D.5, we investigate a practical setting where the observed \mathbf{X} distribution deviates from the population, and discuss how to use the rebalancing method to get trustworthy estimates.

D.1. Simulation Setup

Throughout Sections D.2 and D.3, we assume that the link function G is correctly specified. Consistent with our empirical study in Section 4, we use the *Generalized Sigmoid Form II*, i.e., for each data point i , we have

$$y_i = \frac{\theta_{m+1}^*}{1 + \exp(\theta_0^*(\mathbf{x}_i) + \theta_1^*(\mathbf{x}_i)t_{i1} + \theta_2^*(\mathbf{x}_i)t_{i2} + \dots + \theta_m^*(\mathbf{x}_i)t_{im})} + \epsilon_i, \quad (42)$$

where ϵ_i is the *i.i.d.* random noise with zero mean.

We assume there are m concurrent field experiments, each with a binary treatment. We sample θ_{m+1}^* from the uniform distribution $\mathcal{U}(10, 20)$ throughout, while generating $\theta_j(\mathbf{x})$, $j = 0, 1, \dots, m$, differently in different subsequent sections. We generate data points $\mathbf{z}_i = (y_i, \mathbf{x}'_i, \mathbf{t}'_i)'$ as *i.i.d.* copies of the random vector $\mathbf{Z} = (Y, \mathbf{X}', \mathbf{T}')'$. The random perturbation ϵ follows a uniform distribution $\mathcal{U}(-0.05, 0.05)$. Without loss of generality, we generate covariates \mathbf{x} as follows: (1) the dimension of covariates \mathbf{X} satisfies $d_{\mathbf{X}} = 10$; (2) the

Table A3 Detailed Results of Benchmark Estimators

Treatment Combination	Ground-Truth ATE (1)	Estimator (2)	Estimated				
			ATE (3)	CD (4)	APE (5)	SE (6)	AE (7)
(0, 0, 1)	1.091%**	LA	1.091%	1	0.00%	0.000	0.000
		LR	1.329%	1	21.79%	5.657	2.379
		PDL	1.247%	1	14.24%	2.417	1.555
		SDL	1.179%	1	8.00%	0.763	0.873
(0, 1, 0)	-0.267%	LA	-0.267%	1	NA	0.000	0.000
		LR	-0.013%	1	NA	6.460	2.542
		PDL	-0.036%	0	NA	5.353	2.314
		SDL	-0.072%	0	NA	3.792	1.947
(1, 0, 0)	0.758%*	LA	0.758%	1	0.00%	0.000	0.000
		LR	1.079%	1	42.29%	10.028	3.206
		PDL	1.043%	1	37.60%	8.126	2.851
		SDL	0.978%	1	28.95%	4.816	2.195
(1, 1, 1)	2.121%****	LA	2.121%	1	0.00%	0.000	0.000
		LR	2.395%	1	12.95%	7.539	2.746
		PDL	2.326%	1	9.67%	4.209	2.052
		SDL	2.040%	1	3.78%	0.642	0.801
(1, 1, 0)	0.689%	LA	0.491%	1	NA	3.902	1.975
		LR	1.066%	0	NA	14.233	3.773
		PDL	1.030%	0	NA	11.625	3.410
		SDL	0.902%	0	NA	4.543	2.132
(1, 0, 1)	2.299%****	LA	1.850%	1	19.56%	20.229	4.498
		LR	2.408%	1	4.72%	1.178	1.085
		PDL	2.333%	1	1.46%	0.112	0.336
		SDL	2.148%	1	6.59%	2.297	1.516
(0, 1, 1)	1.387%***	LA	0.824%	0	40.58%	31.670	5.628
		LR	1.316%	1	5.08%	0.495	0.704
		PDL	1.217%	1	12.26%	2.890	1.700
		SDL	1.070%	1	22.84%	10.030	3.167

Notes: The calculation of APE, SE, and AE is based on the scaled outcome variable (see column (1) of this table). SE is scaled by multiplying a constant. AE is scaled by multiplying another constant. The significance levels are encoded as * $p<0.05$; ** $p<0.01$; *** $p<0.001$; **** $p<0.0001$.

different components of \mathbf{X} are *i.i.d.* following the uniform distribution $\mathcal{U}(0, 1)$. We remark that larger random perturbations, higher dimensional covariates, and/or more complicated joint distributions of \mathbf{X} can be easily incorporated into the simulation. We adopt the current setting for ease of model training and efficiency of experiments.

To ensure that the identifiability and sufficient curvature conditions (i.e., Assumption 4) are met, we adopt the following treatment assignment mechanism $\nu(\cdot | \cdot)$. We assume the independence between \mathbf{X} and \mathbf{T} , whose distribution we denote as $\nu(\mathbf{t}) = \mathbb{P}[\mathbf{T} = \mathbf{t}]$. Furthermore, in the training stage, with equal probability, we randomly assign each experimental unit to one of $m + 2$ different treatment combinations with equal probability, i.e., $\nu(\mathbf{t}) = 1/(m + 2)$ if $\mathbf{t} \in \{\mathbf{t} \in \{0, 1\}^m : \sum_{i=1}^m t_i = 0 \text{ or } 1\} \cup \{(1, 1, 0, \dots, 0)'\}$ and $\nu(\mathbf{t}) = 0$ for other treatment combinations. In other words, we assume the *partially observed outcome* setting with only

$m + 2$ observable treatment combinations, while other treatment outcomes are masked for gauging the performance of estimators.

Our neural network structure to estimate $\theta_j^*(\mathbf{x})$, $j = 0, 1, \dots, m$, is prespecified as two-layer perceptrons with ReLU activation functions and 10 nodes in hidden layers in all experiments throughout this section. Because there are $m + 2$ unknown parameters, which linearly scale with the number of experiments m , we generate $500m$ *i.i.d.* experimental data points \mathbf{z}_i for the DNN training, and another independent $500m$ experimental data points for the inference stage. During the DNN training stage, we randomly split this data set with 70% for training and the rest 30% for cross-validation. We adopt the MSE loss function and Adam algorithm (Kingma and Ba 2014). In most experiments, we stop training when the loss on cross-validation data is less than a fixed threshold of 0.3. We empirically tested various thresholds and found that the gain from a smaller stopping threshold is marginal. Thus, we picked this threshold based on the computational efficiency consideration. In the experiments with misspecified link functions and imbalanced data, which we will discuss later, the cross-validation loss tends to increase. Hence, we adjust the threshold accordingly in such cases. We also experiment with popular training strategies such as dropout or regularizing the weights, but the gain is marginal, so we do not include them in this discussion.

At the inference stage, we independently generate a data sample with the same size as the training data. To avoid the rare case where the empirical estimate $\hat{\Lambda}(\mathbf{x})$ is not invertible (e.g., $\theta_i(\mathbf{x}) = \theta_j(\mathbf{x})$ for all $i, j \in \{1, 2, \dots, m\}$), we add a small regularization to $\hat{\Lambda}(\mathbf{x})$ so that $(\hat{\Lambda}(\mathbf{x}) + 0.0005\mathbf{I}_{d_{m+2}})^{-1}$ is well-defined. Similar regularization or trimming techniques based on the propensity score are quite common in practice for numerical stability.

To calculate the true ATE over the population, we use the sample average of 2,000 independent samples for each treatment combination, and use the standard t -test with a significance level of 0.05 to determine whether the ATE of an experiment combination is statistically significant. To derive statistical metrics such as confidence intervals, we replicate all experiments 200 times.

D.2. Validation of the DeDL Estimator under Large m

In this section, we aim at empirically validating the theoretical results in Section 3 and further demonstrating the superior performance of our DeDL estimator in practice. Such experiments are necessary due to the gaps between theory and practice. In particular, there are two potential inconsistencies between the underlying theory (e.g., see Section 3) and practical settings. First, the key theoretical result Proposition 1 is proved under the assumption that one obtains the estimator $\hat{\theta}(\mathbf{x})$ by (almost) minimizing the empirical loss. However, the loss of DNN is difficult to optimize globally, especially given our novel structured architecture with a model layer. Second, the theoretical DNN width $O(n^{d_{\mathbf{x}}/2(p+d_{\mathbf{x}})} \log^2 n)$ and depth $O(\log n)$ required in Proposition 1 are clearly too large for practical applications. Practitioners often fine-tune these hyperparameters at a much smaller scale. Given these practical issues, we find it necessary to conduct synthetic experiments to demonstrate the performance of DeDL in general settings.

To generate the functions $\boldsymbol{\theta}^*(\mathbf{x})$ in this subsection, we first define the coefficient matrix $\mathbf{A} \in \mathbb{R}^{(m+1) \times d_{\mathbf{x}}}$ with each component independently drawn from the uniform distribution $\mathcal{U}(-0.5, 0.5)$ and write the j the row of \mathbf{A} as row vector $\mathbf{A}_{[j]}$. Then, we let $\theta_j^*(\mathbf{x}) = (\mathbf{A}_{[j+1]}\mathbf{x})^3$, $j = 0, 2, \dots, m$. As mentioned, the parameter θ_{m+1}^* is

randomly generated from the uniform distribution $\mathcal{U}(10, 20)$. To facilitate numerical experiments, we choose a relatively simple structure for the neural network, as discussed previously. However, the more complex function $\theta^*(\mathbf{x})$ can be readily incorporated and tested in our synthetic experiments, and DNN training-related hyperparameters (e.g., width, depth, and training algorithm) should be fine-tuned to accommodate such cases.

We summarize the main result here while deferring more details into Appendices D.2.1 and D.2.2. There, we report complete simulation results in Table A4 and Table A5, along with more observations made. In Figure A1(a), we evaluate the performance of different estimators including LA, LR, PDL, SDL, and DeDL with varying numbers of A/B tests, i.e., $m \in \{4, 6, 8, 10\}$. We train these estimators with partially observed outcomes, as discussed above. We note that the performance of MAE is highly correlated with other performance metrics, so to keep the discussion simple, we mainly report and visualize the performance comparison under MAE along with the 95% confidence interval, shown in Figure A1. We compute the confidence interval using the 200 instances for each parameter combination. One can observe from panel (a) that DeDL has the best performance under all $m \in \{4, 6, 8, 10\}$. Increased values of m lead to quick degradation of the performance of LA and LR. Such simple models relying on linear extrapolation are unable to capture the rich treatment effects, but the performance of SDL and DeDL are relatively stable.

To highlight the issue that PDL can easily overfit the observed data and result in large biases in unobserved treatments, we conduct a more detailed synthetic experiment focusing on PDLs with different network size specifications and partially/fully observed data. Figure A1(b) displays the performance of different variants of PDL estimators under different m values. For better visualization, we report the performance of PDL with a different scale on the y-axis than that in Figure A1(a). Among these different PDL estimators, we use subscripts s (small) and l (large) to represent different widths of neural nets, with 10 (small) and 40 (large) hidden nodes for all three hidden layers, respectively. All DNNs have three linear layers followed by ReLU activation layers. The subscripts p and a represent the training samples generated from *partially observed treatments* and *all* treatments, respectively. For a fair comparison to SDL and DeDL, we focus on PDL_s_p with partially observed treatments and a similar DNN size to SDL. We observe that, with partially observed combinations, both PDL_s_p and PDL_l_p perform poorly, mainly driven by the bad performance under the unobserved treatments. Increased network size does not help much. However, when we incorporate data from all treatment combinations for training, PDL_s_a and PDL_l_a obtain comparable performance with the best estimator DeDL. Therefore, a structured DNN model allows us to capture the interaction of experiments in practical settings, while the model flexibility in PDL helps only in unrealistic scenarios where a large proportion of experimental combinations are observed. We refer interested readers to Appendix ?? for more detailed comparisons, discussions, and additional simulations of incorporating regularizers into PDL.

Because all performance metrics show similar patterns under different m values, we maintain $m = 4$ for computation efficiency in subsequent sections. Also, due to the inferior performance of PDL with partially observed treatments, we do not report its performance in the following sections.

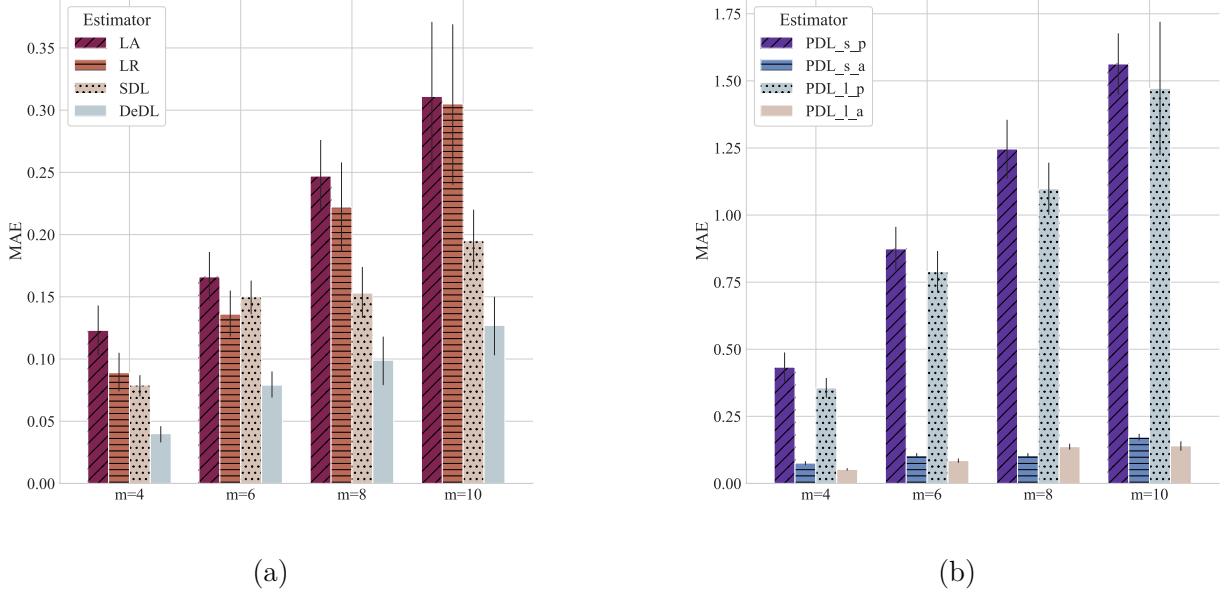


Figure A1 MAE comparison among estimators under the increasing number of experiments m values. Panel (a) shows the performance of LA, LR, SDL, and DeDL. Panel (b) presents the performance of PDLs with different network size specifications and partially/fully observed data. Subscript s represents a small 10-width DNN while l represents a large 40-width DNN. Subscript p represents that the training set contains partially observed treatment while a represents that the training set contains all 2^m treatments.

D.2.1. Detailed Results of the Comparison Among Estimators. We document the complete simulation results of Appendix D.2 in Table A4 and Table A5. We evaluate the performance of the estimators under $m \in \{4, 6, 8, 10\}$ A/B tests and report the estimation results in Panels A to D. The first column “Estimator” describes which estimator is tested. The second column “CDR” shows the proportion of treatment combinations whose estimated ATE significance levels and signs are consistent with the ground truth. The third column “MAPE” gives the mean absolute percentage error of ATE estimates over all treatment combinations whose real average treatment effects are significant. In other words, we rule out insignificant treatment combinations when calculating MAPE. Otherwise, those insignificant treatment effects would result in a close-to-zero value in the denominators to calculate APE, resulting in an undesired metric. Similarly, column “MSE” and column “MAE” represent squared error and absolute error, respectively. In these two columns, we do not exclude those insignificant combinations. Indeed, unlike MAPE, close-to-zero treatment effects do not cause a problem for MSE and MAE. As a result, MAE and MSE over all combinations can supplement MAPE. However, to better understand the scales of MSE and MAE errors, we also report in the table notes 95% confidence intervals of average absolute treatment effects over all combinations. Finally, the column “BTI” presents the results on best treatment identification. For each replication of the experiment, we verify whether different methods can successfully identify the best treatment combination. The value in this column shows the proportion of replications in which the optimal combination is correctly identified.

We observe the following consistent patterns across all panels in Table A4. First, DeDL outperforms LA, LR, PDL_{s_p}, and SDL under all metrics, which validates our theory and provides strong evidence for the advantage of our method. In particular, DeDL increases the success rate of both CDR and BTI, and decreases MAPE, MSE, and MAE compared to SDL by a significant margin. This demonstrates the value of debiasing

Table A4 Learning Estimator Validation Result

Panel A: Comparison of Different Estimators Under $m = 4$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	92.40% (91.02%, 93.78%)	22.64% (20.10%, 25.17%)	0.070 (0.023, 0.118)	0.123 (0.103, 0.143)	79.5% (73.9%, 85.1%)
LR	95.34% (94.46%, 96.23%)	15.30% (13.56%, 17.04%)	0.038 (0.0179, 0.0582)	0.089 (0.074, 0.105)	82.0% (76.6%, 87.3%)
PDL_s_p	88.59% (87.02%, 90.17%)	68.83% (60.76%, 76.89%)	0.867 (0.572, 1.163)	0.433 (0.378, 0.488)	56.0% (49.1%, 62.9%)
SDL	95.12% (94.17%, 96.08%)	16.12% (14.57%, 17.66%)	0.018 (0.011, 0.024)	0.079 (0.070, 0.087)	90.5% (86.4%, 94.6%)
DeDL	97.53% (96.90%, 98.15%)	7.20% (6.45%, 7.95%)	0.008 (0.004, 0.012)	0.040 (0.033, 0.046)	93.5% (90.1%, 96.9%)
Panel B: Comparison of Different Estimators Under $m = 6$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	92.05% (90.70%, 93.40%)	27.35% (23.81%, 30.89%)	0.080 (0.056, 0.105)	0.166 (0.147, 0.186)	73.0% (66.8%, 79.2%)
LR	94.88% (94.10%, 95.67%)	18.42% (16.61%, 20.24%)	0.073 (0.038, 0.108)	0.136 (0.118, 0.155)	75.5% (69.4%, 81.6%)
PDL_s_p	84.23% (82.54%, 85.93%)	111.70% (98.07%, 125.33%)	2.172 (1.736, 2.607)	0.874 (0.793, 0.956)	28.0% (21.7%, 34.3%)
SDL	93.07% (92.17%, 93.96%)	27.22% (25.16%, 29.29%)	0.051 (0.037, 0.064)	0.150 (0.136, 0.163)	68.0% (61.5%, 74.5%)
DeDL	95.28% (94.61%, 95.96%)	13.33% (11.99%, 14.68%)	0.023 (0.012, 0.033)	0.079 (0.069, 0.090)	87.0% (82.3%, 91.7%)
Panel C: Comparison of Different Estimators Under $m = 8$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	93.83% (92.74%, 94.92%)	28.37% (24.98%, 31.75%)	0.196 (0.135, 0.256)	0.247 (0.218, 0.276)	65.0% (58.3%, 71.7%)
LR	95.51% (94.87%, 96.14%)	22.49% (20.03%, 24.96%)	0.209 (0.128, 0.289)	0.222 (0.186, 0.258)	75.6% (69.2%, 82.0%)
PDL_s_p	78.89% (77.01%, 80.77%)	140.60% (128.28%, 152.91%)	3.825 (3.059, 4.591)	1.246 (1.136, 1.355)	11.5% (7.0%, 16.0%)
SDL	94.89% (94.25%, 95.53%)	22.18% (20.20%, 24.15%)	0.072 (0.025, 0.120)	0.153 (0.133, 0.174)	67.5% (61.0%, 74.0%)
DeDL	95.53% (94.92%, 96.15%)	13.49% (11.82%, 15.15%)	0.049 (0.007, 0.092)	0.099 (0.079, 0.118)	82.0% (76.6%, 87.4%)
Panel D: Comparison of Different Estimators Under $m = 10$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	95.01% (93.96%, 96.05%)	31.54% (24.23%, 38.84%)	0.326 (0.152, 0.500)	0.311 (0.251, 0.371)	58.8% (47.7%, 69.8%)
LR	95.51% (94.87%, 96.14%)	25.28% (21.73%, 28.83%)	0.359 (0.146, 0.572)	0.305 (0.240, 0.369)	62.0% (52.3%, 71.7%)
PDL_s_p	77.27% (75.50%, 79.03%)	163.85% (151.47%, 176.22%)	5.360 (4.527, 6.193)	1.563 (1.448, 1.677)	5.0% (2.0%, 8.0%)
SDL	94.27% (93.21%, 95.33%)	26.37% (22.97%, 29.78%)	0.077 (0.054, 0.100)	0.195 (0.171, 0.220)	56.2% (45.1%, 67.4%)
DeDL	94.54% (93.48%, 95.60%)	16.59% (13.39%, 19.78%)	0.046 (0.026, 0.067)	0.127 (0.103, 0.150)	78.8% (69.6%, 87.9%)

Notes: All experiments are replicated 200 times, with 95% CIs reported in parentheses. 95% CIs of average absolute treatment effects are (0.68, 1.06), (1.09, 1.30), (1.45, 1.71), and (1.54, 2.03), respectively, in Panel A, Panel B, Panel C, and Panel D.

Table A5 Performance of PDL Estimators

Panel A: Comparison of Different Estimators Under $m = 4$					
Estimator	CDR	MAPE	MSE	MAE	BTI
PDL_s_p	88.59% (87.02%, 90.17%)	68.83% (60.76%, 76.89%)	0.867 (0.572, 1.163)	0.433 (0.378, 0.488)	56.0% (49.1%, 62.9%)
PDL_s_a	95.16% (94.27%, 96.04%)	16.01% (14.49%, 17.53%)	0.012 (0.009, 0.016)	0.075 (0.068, 0.083)	78.0% (72.2%, 83.8%)
PDL_l_p	89.12% (87.61%, 90.64%)	60.15% (53.09%, 67.20%)	0.493 (0.375, 0.610)	0.356 (0.319, 0.393)	56.0% (49.1%, 62.9%)
PDL_l_a	96.34% (95.57%, 97.12%)	12.55% (11.20%, 13.89%)	0.006 (0.004, 0.007)	0.052 (0.048, 0.057)	86.5% (81.7%, 91.3%)
Panel B: Comparison of Different Estimators Under $m = 6$					
Estimator	CDR	MAPE	MSE	MAE	BTI
PDL_s_p	84.23% (82.54%, 85.93%)	111.70% (98.07%, 125.33%)	2.172 (1.736, 2.607)	0.874 (0.793, 0.956)	28.0% (21.7%, 34.3%)
PDL_s_a	95.04% (94.43%, 95.65%)	20.03% (18.39%, 21.67%)	0.021 (0.017, 0.024)	0.104 (0.096, 0.113)	70.5% (64.1%, 76.9%)
PDL_l_p	83.87% (82.33%, 85.40%)	111.37% (99.17%, 123.57%)	1.748 (1.357, 2.138)	0.789 (0.712, 0.866)	25.0% (18.9%, 31.1%)
PDL_l_a	95.73% (95.05%, 96.42%)	16.91% (14.93%, 18.89%)	0.015 (0.012, 0.019)	0.085 (0.077, 0.093)	68.0% (61.5%, 74.5%)
Panel C: Comparison of Different Estimators Under $m = 8$					
Estimator	CDR	MAPE	MSE	MAE	BTI
PDL_s_p	78.89% (77.01%, 80.77%)	140.60% (128.28%, 152.91%)	3.825 (3.059, 4.591)	1.246 (1.136, 1.355)	11.5% (7.0%, 16.0%)
PDL_s_a	95.04% (94.43%, 95.65%)	20.03% (18.39%, 21.67%)	0.021 (0.017, 0.024)	0.104 (0.096, 0.113)	70.5% (64.1%, 76.9%)
PDL_l_p	80.80% (79.15%, 82.44%)	125.38% (114.46%, 136.29%)	3.059 (2.347, 3.770)	1.098 (1.001, 1.195)	10.0% (5.8%, 14.2%)
PDL_l_a	95.43% (94.90%, 95.95%)	20.24% (18.36%, 22.13%)	0.035 (0.028, 0.041)	0.137 (0.126, 0.148)	57.5% (50.6%, 64.4%)
Panel D: Comparison of Different Estimators Under $m = 10$					
Estimator	CDR	MAPE	MSE	MAE	BTI
PDL_s_p	77.27% (75.50%, 79.03%)	163.85% (151.47%, 176.22%)	5.360 (4.527, 6.193)	1.563 (1.448, 1.677)	5.0% (2.0%, 8.0%)
PDL_s_a	95.14% (94.69%, 95.59%)	24.87% (22.64%, 27.11%)	0.052 (0.044, 0.061)	0.173 (0.160, 0.185)	43.0% (36.1%, 49.9%)
PDL_l_p	76.03% (73.04%, 79.01%)	159.58% (124.10%, 195.06%)	5.436 (3.533, 7.339)	1.472 (1.224, 1.720)	6.0% (1.2%, 10.8%)
PDL_l_a	95.67% (94.82%, 96.53%)	19.66% (17.00%, 22.31%)	0.033 (0.025, 0.041)	0.140 (0.122, 0.157)	56.0% (41.7%, 70.3%)

Note: All experiments are replicated 200 times, with 95% CIs reported in parentheses.

in the influence function (5). Second, LA and LR perform worse than SDL in general, demonstrating the advantage of neural networks over linear methods, although SDL may still be asymptotically biased.

Comparing across different panels, we observe that the performance of all estimators becomes worse when m grows larger. In particular, the performance of BTI worsens quickly due to the exponentially increased number of combinations. Even so, DeDL can still successfully identify the best treatment among the 1,024 combinations with a relatively high probability of 78.8% when $m = 10$.

We also remark on the underlying mechanisms of these degenerating performances. On one hand, intuitively, due to the sigmoid link function setup, when the number of field experiments gets larger, the treatment effect becomes more nonlinear. It creates difficulty for the LA estimator. Therefore, the performance of LA worsens due to its inherent lack of model richness. On the other hand, with the correct link function specification, one may expect that the performance of SDL and DeDL should be relatively stable because of their strong modeling power. However, since we fix the same DNN structure with a constant number of 10 hidden nodes across all experiments for a fair comparison with LA, the model complexity of neural networks is limited by design. We have verified that increasing the number of hidden nodes for $m = 6, 8, 10$ helps achieve similar performance to that of $m = 4$. Indeed, the challenge of estimation and inference from an increased m can be mitigated by increasing the size and complexity of the DNN. Therefore, without loss of generality, for all experiments in this section, we keep $m = 4$ for computational efficiency.

D.2.2. A Closer Look at PDL. We report the performance of the PDL estimators in a separate table, i.e., Table A5, to provide an anatomy of their bad performance. The subscripts s and l represent different widths of neural nets, with 10 and 40 hidden nodes for hidden layers, resulting in the DNN structure $(d_x + d_t + 1)\text{-ReLU-10-ReLU-10-ReLU-1}$ and $(d_x + d_t + 1)\text{-40-ReLU-40-ReLU-40-ReLU-1}$, respectively. The subscripts p and a represent the training samples generated from partially observed treatments or all treatments. For a fair comparison to estimators in Table A4, we should focus on PDL_s_p with partially observed treatments and similar DNN size as SDL. The training stopping criterion is set the same as SDL with 0.3 validation loss threshold.

With partially observed combinations, both PDL_s_p and PDL_l_p perform worse than all other estimators under all metrics. This is due to the bad performance of the out-of-sample test under the unobserved treatments, rather than the bad approximation ability of DNN. When we increase the DNN width from 10 to 40, the performance only slightly increases. Also, PDL indeed has much better in-sample tests. Because when we incorporate data from all treatment combinations in the training, the resulting estimators PDL_s_a and PDL_l_a have comparable performance with the best estimator DeDL.

It is within our expectation that PDL with partially observed treatments has such bad performance because it can only access the data generated by base treatment level, m single experiment data, and only one treatment level with interaction $(1, 1, 0, \dots, 0)$. It means that it is almost impossible for PDL to learn interaction between different experiments. Unlike LR and SDL, which impose parametric structures of experimental interaction, PDL aims only at increasing the performance of in-sample outcome prediction, totally ignoring the out-of-sample performance. One may argue that this is due to the over-fitting of PDL.

To explore in more detail why PDL performs badly, we investigate further. To simplify the discussion, we conduct an extra synthetic experiment with $m = 3$; we report the result in Figure A2. Each point in the scatter plots represents the values of real ATE (x-axis) and predicted ATE (y-axis). From left to right, we visualize the performance of LR, PDL-base (i.e., PDL_s_p), PDL-dropout (i.e., PDL_s_p with $p = 0.1$ dropout regularizer after each activation layer), and PDL-L1 (i.e., PDL_s_p with L_1 regularization loss over DNN parameters with fine-tuned weight 0.05).

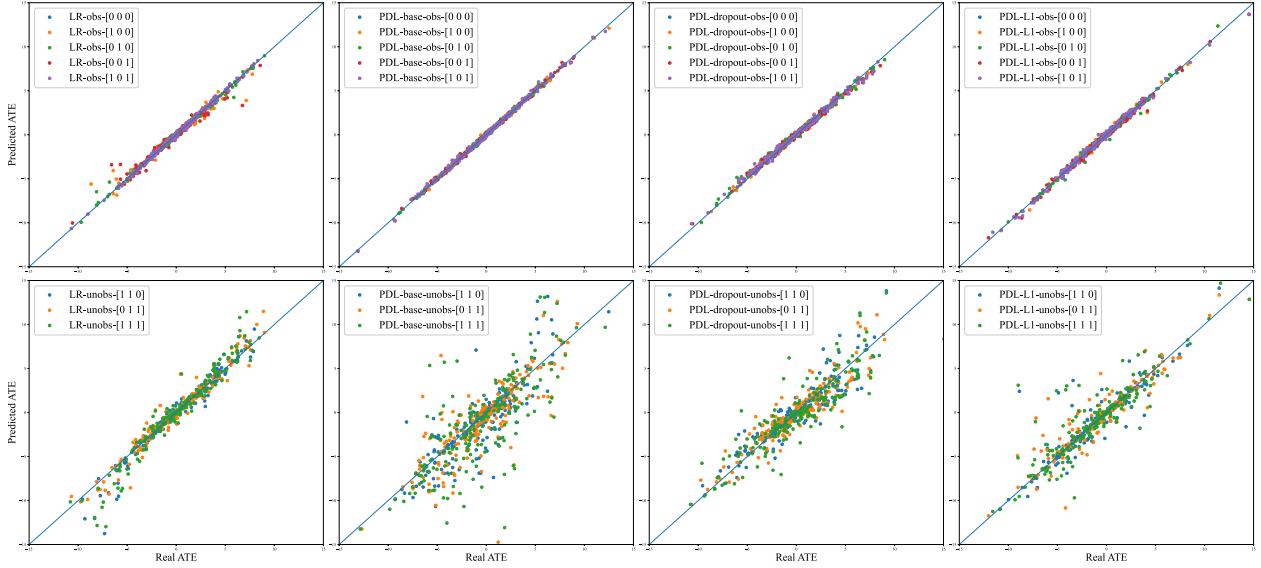
Table A6 Performance of PDL With Regularizers

Estimator	CDR	MAPE	MSE	MAE	BTI
LA	93.56% (92.06%, 95.07%)	27.44% (22.60%, 32.29%)	0.705 (0.504, 0.905)	0.408 (0.359, 0.457)	89.0% (84.6%, 93.4%)
LR	95.50% (94.37%, 96.63%)	26.01% (20.56%, 31.45%)	0.654 (0.449, 0.859)	0.389 (0.335, 0.443)	91.5% (87.6%, 95.4%)
PDL-base	91.62% (90.22%, 93.03%)	52.68% (44.27%, 61.09%)	2.930 (2.104, 3.756)	0.762 (0.675, 0.849)	73.0% (66.8%, 79.2%)
PDL-dropout	90.81% (89.17%, 92.46%)	35.74% (31.43%, 40.04%)	0.967 (0.739, 1.195)	0.504 (0.453, 0.555)	63.5% (56.8%, 70.2%)
PDL-L1	92.00% (90.25%, 93.75%)	33.64% (27.96%, 39.32%)	1.274 (0.513, 2.035)	0.426 (0.354, 0.498)	84.5% (79.4%, 89.6%)

Note: All experiments are replicated 200 times under $m = 3$, with 95% CIs reported in parentheses.

Four subfigures in the upper panel show the in-sample performance, while lower panel subfigures show the performance under unobserved treatment combinations. Notice that PDL-base has better in-sample performance than LR, as the data points are more concentrated around true line $y = x$, but PDL-base has the worst out-of-sample performance. Although LR assumes linear extrapolation of treatment effects, in our example, most scatter points in the out-of-sample test are still well concentrated around $y = x$ except clear patterns of overestimate when absolute ATE increases. When applying regularizations, PDL-dropout and PDL-L1 have deteriorated in-sample performance with points less concentrated around the true line while improving out-of-sample performance.

Generally, we do not know the out-of-sample ground-truth ATE in practice, which makes it difficult to guide the selection of regularizers. We also try other regularizers, such as early stopping, L_2 parameter weights, and smaller network sizes, which, however, still perform badly comparable to LR. Due to the bad performance of PDL with partially observed treatments, we do not report it in the following simulations.

**Figure A2 LR and PDL Estimator Comparison**

D.3. Robustness to the DNN Convergence Rate

As pointed out by the semiparametric estimation literature (e.g., Chernozhukov et al. 2018), in practice, it is sometimes questionable whether DNN estimators $\hat{\theta}(\cdot)$ can achieve the $o(n^{-1/4})$ convergence rate required for the inference stage. To systematically illustrate the performance of the debiasing technique in learning the treatment effects, we artificially control the biases in DNN estimators to evaluate the impact of such biases in the second-stage inference.

Specifically, we use an approach with a similar spirit to Chernozhukov et al. (2018) by constructing $\hat{\theta}(\mathbf{x})$ with manually controlled biases. First, parameter functions are defined as $\theta_j^*(\mathbf{x}) = \mathbf{A}_{[j]}\mathbf{x}$, for $j = 0, 1, \dots, m$, where each component in the coefficient matrix $\mathbf{A} \in \mathbb{R}^{(m+1) \times d_{\mathbf{x}}}$ is generated under independent and uniform distribution $\mathcal{U}(-0.5, 0.5)$. The parameter θ_{m+1}^* is randomly generated from the uniform distribution $\mathcal{U}(10, 20)$. Next, instead of training a DNN for estimation, we manually set the biased estimator $\hat{\theta}_j(\mathbf{x}) = (1 + err_j)\theta_j^*(\mathbf{x})$, $j = 0, 1, \dots, m+1$, where all err_j , $j = 0, 1, \dots, m+1$ terms independently follow the uniform distribution $\mathcal{U}(-\delta, \delta)$. We set different levels of the bias range coefficient $\delta \in \{0.1, 0.2, 0.3\}$ to investigate the effectiveness of DeDL with different levels of biases of the estimators $\hat{\theta}$. In the following discussion, we focus on the MAE performance metric. We refer interested readers to Appendix ?? for a more detailed comparison and discussion.

As documented in Figure A3, DeDL has much smaller MAEs than SDL, implying a significant performance improvement with adding debias term. In all settings, DeDL performs better than LA and LR despite that SDL may generate MAEs much larger than LA and LR, in particular, when $\delta = 0.3$. This shows that the DeDL estimator is fairly robust, with moderate biases in $\hat{\theta}$. However, we also point out that when the bias from training is large, debiasing may not improve the performance, as we illustrated in Figure 7 in Section 4. In our simulation, we have a similar observation that, when δ increases to 1.0, DeDL may perform worse than SDL under the MAPE metric. In such cases, it is critical to improve DeDL by training a better DNN model, for example, through using a larger network. For completeness, we also document the detaild results in Table A7, the columns of which are the same as Table A4. Insights similar to above can be inferred from the table.

D.4. Link Function Misspecification

In this subsection, we first investigate how a misspecified link function impairs the effectiveness of our debiased estimator. Indeed, a key assumption of our framework is that the link function G is correctly specified. In practice, however, it can be challenging to select the best link function. On the positive side, as we will discuss, one may test the efficacy of the link function by checking the cross-validation errors in the DNN training stage. For this subsection, we adjust our true DGP as,

$$y_i = \frac{\theta_{m+1}^*}{1 + \exp(\theta_0^*(\mathbf{x}_i) + \theta_1^*(\mathbf{x}_i)t_{i1} + \dots + \theta_m^*(\mathbf{x}_i)t_{im})} + \gamma(\beta_0^*(\mathbf{x}_i) + \beta_1^*(\mathbf{x}_i)t_{i1} + \dots + \beta_m^*(\mathbf{x}_i)t_{im}) + \epsilon_i, \quad (43)$$

where parameter functions are defined as $\theta_j^*(\mathbf{x}) = \mathbf{A}_{[j]}\mathbf{x}$, and $\beta_j^*(\mathbf{x}) = \mathbf{B}_{[j]}\mathbf{x}$, for $j = 0, 1, \dots, m$. The parameter $\gamma \geq 0$ captures the extent to which the true link function deviates from the *Generalized Sigmoid Form II*, which we still adopt in our DeDL framework for estimating the treatment effect. The larger the γ , the more misspecified our model is. We generate both the coefficient matrices $\mathbf{A} \in \mathbb{R}^{(m+1) \times d_{\mathbf{x}}}$ and $\mathbf{B} \in \mathbb{R}^{(m+1) \times d_{\mathbf{x}}}$ with

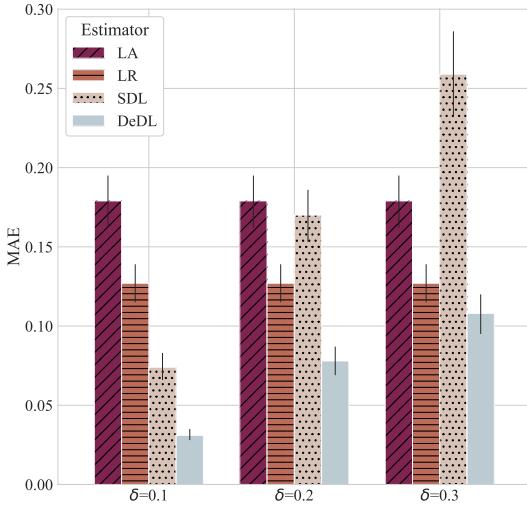


Figure A3 MAE Comparison Under Decreasing DNN Convergence Rate

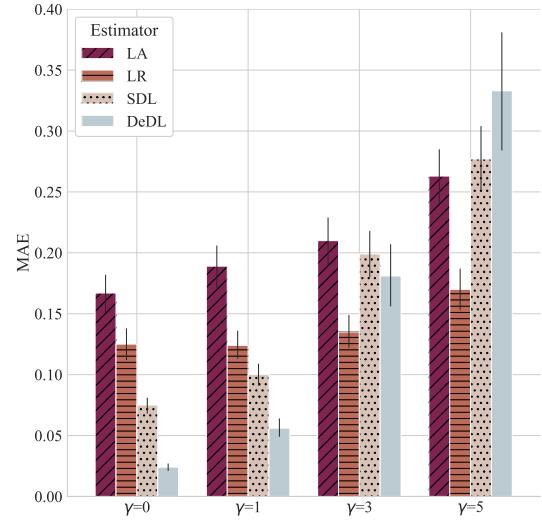


Figure A4 MAE Comparison Under Link Function Misspecification

Table A7 DNN Convergence Violation Result

Panel A: Performance of Benchmarks					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	94.56% (93.69%, 95.44%)	19.93% (17.94%, 21.93%)	0.085 (0.066, 0.104)	0.179 (0.162, 0.195)	92.0% (88.2%, 95.8%)
LR	96.88% (96.21%, 97.54%)	13.59% (12.11%, 15.08%)	0.046 (0.036, 0.056)	0.127 (0.115, 0.139)	94.0% (90.7%, 97.3%)
Panel B: Comparison of Different Estimators Under $\delta = 0.1$					
Estimator	CDR	MAPE	MSE	MAE	BTI
SDL	96.97% (96.32%, 97.62%)	4.98% (4.58%, 5.38%)	0.013 (0.010, 0.015)	0.074 (0.066, 0.083)	92.5% (88.8%, 96.2%)
DeDL	98.56% (98.12%, 99.01%)	5.04% (4.32%, 5.76%)	0.003 (0.002, 0.003)	0.031 (0.028, 0.035)	97.0% (94.6%, 99.4%)
Panel C: Comparison of Different Estimators Under $\delta = 0.2$					
Estimator	CDR	MAPE	MSE	MAE	BTI
SDL	94.19% (93.19%, 95.19%)	11.06% (10.28%, 11.83%)	0.061 (0.049, 0.072)	0.170 (0.153, 0.186)	85.5% (80.6%, 90.4%)
DeDL	96.31% (95.50%, 97.12%)	11.10% (9.46%, 12.74%)	0.016 (0.012, 0.020)	0.078 (0.069, 0.087)	95.5% (92.6%, 98.4%)
Panel D: Comparison of Different Estimators Under $\delta = 0.3$					
Estimator	CDR	MAPE	MSE	MAE	BTI
SDL	93.44% (92.48%, 94.39%)	15.53% (14.33%, 16.73%)	0.145 (0.118, 0.173)	0.259 (0.232, 0.286)	80.0% (74.4%, 85.6%)
DeDL	94.63% (93.74%, 95.51%)	12.78% (11.05%, 14.50%)	0.032 (0.024, 0.039)	0.108 (0.095, 0.120)	92.5% (88.8%, 96.2%)

Notes: All simulations are replicated 200 times, with 95% CIs reported in parentheses. 95% CIs of average absolute treatment effect are (1.45, 1.65) in all panels.

independent entries, and each component follows the uniform distribution $\mathcal{U}(-0.5, 0.5)$. The parameter θ_{m+1}^* is again randomly generated from the uniform distribution $\mathcal{U}(10, 20)$.

We run synthetic experiments under different levels of model misspecification $\gamma \in \{0, 1, 3, 5\}$. To illustrate how the link function is skewed, we plot in Figure A5 the histograms of outcome y in (43). The x -axis represents the experimental outcome y among the population in all 2^m treatment combinations with equal probability. Observe that when γ is large, e.g., $\gamma = 5$, the bias from using the sigmoid link function to approximate the experimental outcome is likely to be large. Because the DGP defined by (43) may be too restrictive in the sense of ATE representation, further in Appendix D.4.2, we conduct another synthetic experiment where the true DGP has explicit higher-order interaction terms. The results show that debiasing through Neyman orthogonality is fairly robust under such model misspecification.

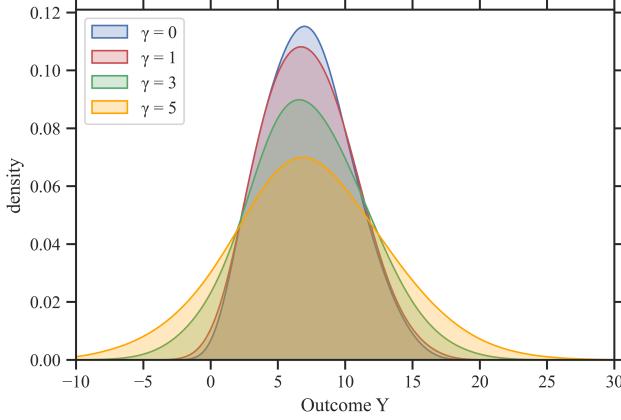


Figure A5 Experimental Outcome of Misspecified Model

Defferring reporting the complete results in Appendix D.4.1, here we emphasize the main insights. We first plot the comparison between different methods under model misspecification in Figure A4. This comparison reveals that when the link function is more specified (i.e., larger γ), **SDL** and **DeDL** estimators perform substantially worse, whereas the performances of the **LA** and **LR** estimators are relatively stable. Indeed, the **LA** and **LR** estimators face no significant increases in MAPE as γ gets larger. Also, when γ is increased to 5, **DeDL** performs worse than **SDL**, implying that debiasing via Neyman orthogonality hurts the performance when the link function is not correctly specified.

In practice, however, it is difficult, if not impossible, to verify the true link function. Fortunately, one may detect link function misspecification through large training errors. In other words, we can resort to checking the cross-validation errors in the DNN training stage. To shed light on this point, we report the training errors from our experiments. Specifically, we compare the errors induced by pure DNN without a sigmoid link function (a three-layer perceptron in our case) and structured DNN (a two-layer perceptron followed by a link function layer), respectively. Pure DNN takes treatment level t together with covariates \mathbf{x} as inputs to the first linear layer. In contrast, the structured DNN takes only covariates \mathbf{x} to the first linear layer and uses treatment level t in the final link function layer.

For a fair comparison, we set the pure DNN structure with a similar width and depth, i.e., $(d_{\mathbf{x}} + d_t + 1) - 10 - \text{ReLU} - 10 - \text{ReLU} - 10 - \text{ReLU} - 1$. Equipped with higher flexibility, pure DNN is generally good at fitting individual responses in the partially observed treatment combinations. Hence, we can use the pure DNN as a

benchmark for the in-sample comparison to check whether the link function is reasonable. Adopting the same Adam algorithm and training samples, we obtain the following 95% \mathcal{CIs} of cross-validation mean squared errors under different misspecification levels: (a) [0.044, 0.053], [0.086, 0.112], [0.116, 0.159], [0.142, 0.197] for pure DNN under $\gamma = 0, 1, 3, 5$ respectively; and (b) [0.012, 0.014], [0.019, 0.026], [0.076, 0.106], [0.215, 0.279] for structured DNN under $\gamma = 0, 1, 3, 5$ respectively. One can observe that when $\gamma \in \{0, 1, 3\}$, the structured DNN has a smaller or comparable in-sample loss than the pure DNN, which indicates the reasonable performance of the generalized sigmoid link function to approximate the outcome variable. However, when the misspecification level $\gamma = 5$, the in-sample error from the generalized sigmoid form grows larger than that of the pure DNN. Correspondingly, in Figure A4, we observe severe performance degradation for both **SDL** and **DeDL** and a significant negative impact from the debiasing term. In this case, we suggest experimenting with a different link function in the first DNN training stage and/or not using the debiased estimator. More generally, as long as the in-sample loss of the structured DNN is on par with that of a pure DNN with similar depth and width, we recommend adopting this structured DNN with debiasing.

D.4.1. Complete Results of the Experiments in Appendix D.4. Table A8 shows the complete simulation results. Comparing across different panels representing different values of γ , one can observe that when the link function gets more misspecified, **SDL** and **DeDL** estimators get worse under all metrics. While the MAPE, MSE, and MAE performances of **LA** and **LR** are not significantly deteriorated, considering the increased absolute treatment effects listed in table notes. When the model is not misspecified or marginally misspecified, i.e., $\gamma \in \{0, 1, 3\}$, **SDL** and **DeDL** work no worse than **LA** in all performance metrics. When $\gamma = 1$, **DeDL** can only marginally improve the performance of **SDL** for all metrics. However, when γ is increased to 3, we can observe that **DeDL** has worse MSE performance than **SDL**. Even worse, when γ is increased to 5, **DeDL** is not better than **SDL** under all performance metrics. In summary, we still recommend using **DeDL** when $\gamma \in \{0, 1, 3\}$.

D.4.2. Experiment on Model Misspecification under Higher-Order Treatment Interactions.

We conduct a new set of synthetic experiments on the model misspecification with explicit higher-order interaction terms. Specifically, the true data-generating process with all higher-order terms is defined as

$$\begin{aligned} y_i = & \theta_0^*(\mathbf{x}_i) + \theta_1^*(\mathbf{x}_i)t_1 + \theta_2^*(\mathbf{x}_i)t_2 + \theta_3^*(\mathbf{x}_i)t_3 + \theta_4^*(\mathbf{x}_i)t_4 \\ & + \theta_5^*(\mathbf{x}_i)t_1t_2 + \theta_6^*(\mathbf{x}_i)t_1t_3 + \theta_7^*(\mathbf{x}_i)t_1t_4 + \theta_8^*(\mathbf{x}_i)t_2t_3 + \theta_9^*(\mathbf{x}_i)t_2t_4 + \theta_{10}^*(\mathbf{x}_i)t_3t_4 \\ & + \theta_{11}^*(\mathbf{x}_i)t_1t_2t_3 + \theta_{12}^*(\mathbf{x}_i)t_2t_3t_4 + \theta_{13}^*(\mathbf{x}_i)t_3t_4t_1 + \theta_{14}^*(\mathbf{x}_i)t_4t_1t_2 \\ & + \theta_{15}^*(\mathbf{x}_i)t_1t_2t_3t_4 + \epsilon_i, \end{aligned} \tag{44}$$

where all parameters $\theta_i^*(\mathbf{x}) = \mathbf{c}'\mathbf{x}$ ($i = 0, 1, \dots, 15$) are linear in the covariate vector. We generate the coefficient $\mathbf{c} \in \mathbb{R}^{d_x}$ with independent entries, and each component follows the uniform distribution $\mathcal{U}(0, 0.5)$. Note that our proposed *Generalized Sigmoid Form II* is largely misspecified compared to the above true DGP. We run synthetic experiments under different levels of model misspecification by incorporating different terms into the DGP. Specifically, we conduct four different tests. In the first test, we add all terms, i.e., 1st, 2nd, 3rd, and 4th-order interaction terms. In the second test, we only put 1st, 2nd, and 3rd-order terms in the true DGP. In the third test, we have 1st and 2nd-order terms, while the last test only has linear terms.

Table A8 Model Misspecification Result

Panel A: Comparison of Different Estimators Under $\gamma = 0$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	94.78% (93.80%, 95.76%)	20.69% (18.30%, 23.08%)	0.074 (0.056, 0.091)	0.167 (0.151, 0.182)	94.5% (91.3%, 97.7%)
LR	97.38% (96.81%, 97.94%)	12.81% (11.53%, 14.08%)	0.043 (0.033, 0.053)	0.125 (0.112, 0.138)	94.5% (91.3%, 97.7%)
SDL	97.71% (97.18%, 98.26%)	10.07% (8.92%, 11.21%)	0.014 (0.010, 0.018)	0.075 (0.068, 0.081)	96.5% (93.9%, 99.1%)
DeDL	99.28% (98.99%, 99.57%)	3.23% (2.68%, 3.78%)	0.002 (0.001, 0.003)	0.024 (0.021, 0.027)	99.0% (97.6%, 100.4%)
Panel B: Comparison of Different Estimators Under $\gamma = 1$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	95.56% (94.77%, 96.35%)	21.76% (19.29%, 24.23%)	0.091 (0.069, 0.112)	0.189 (0.171, 0.206)	87.5% (82.9%, 92.1%)
LR	97.38% (96.79%, 97.96%)	12.85% (11.45%, 14.24%)	0.042 (0.033, 0.052)	0.124 (0.113, 0.136)	88.0% (83.5%, 92.5%)
SDL	97.22% (96.57%, 97.86%)	11.55% (10.40%, 12.70%)	0.027 (0.019, 0.035)	0.100 (0.091, 0.109)	90.5% (86.4%, 94.6%)
DeDL	98.81% (98.33%, 99.29%)	5.67% (4.87%, 6.47%)	0.014 (0.008, 0.020)	0.056 (0.049, 0.064)	92.5% (88.8%, 96.2%)
Panel C: Comparison of Different Estimators Under $\gamma = 3$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	94.62% (93.71%, 95.54%)	18.07% (16.18%, 19.96%)	0.112 (0.075, 0.148)	0.210 (0.190, 0.229)	89.0% (84.6%, 93.4%)
LR	97.09% (96.46%, 97.73%)	12.10% (10.77%, 13.43%)	0.050 (0.036, 0.063)	0.135 (0.122, 0.149)	88.0% (83.5%, 92.5%)
SDL	97.00% (96.33%, 97.67%)	15.10% (13.54%, 16.66%)	0.123 (0.088, 0.159)	0.199 (0.180, 0.218)	90.5% (86.4%, 94.6%)
DeDL	97.50% (96.77%, 98.22%)	13.44% (10.87%, 16.00%)	0.151 (0.096, 0.207)	0.181 (0.156, 0.207)	92.5% (88.8%, 96.2%)
Panel D: Comparison of Different Estimators Under $\gamma = 5$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	95.34% (94.47%, 96.22%)	19.93% (17.12%, 22.74%)	0.155 (0.125, 0.185)	0.263 (0.241, 0.285)	88.0% (83.5%, 92.5%)
LR	96.97% (96.21%, 97.73%)	12.05% (10.28%, 13.81%)	0.077 (0.053, 0.102)	0.170 (0.153, 0.187)	92.5% (88.8%, 96.2%)
SDL	95.90% (95.09%, 96.72%)	17.24% (15.21%, 19.26%)	0.241 (0.183, 0.299)	0.277 (0.251, 0.304)	88.5% (84.0%, 93.0%)
DeDL	95.46% (94.43%, 96.50%)	23.20% (18.31%, 28.09%)	0.559 (0.375, 0.743)	0.333 (0.284, 0.381)	84.5% (79.4%, 89.6%)

Notes: All simulations are replicated 200 times, with 95% CIs reported in parentheses. 95% CIs of average absolute treatment effects are (1.45, 1.65), (1.52, 1.72), (2.06, 2.36), and (2.71, 3.07) respectively in Panel A, Panel B, Panel C, and Panel D.

The complete result are reported in Table A9, where each panel shows a test result with different terms in the DGP as defined above. Because there are no significant differences in the performance metrics BTI and CDR, we only document the MAPE, MSE, and MAE measures. One can observe that DeDL performs best when there exist interactions among the treatment effects of different experiments; see Panel A, B, and C in Table A9. It means that even the link function is misspecified, DeDL can still capture some level of interactions and make a relatively accurate estimation. Comparing DeDL and SDL, Table A9 shows that

DeDL improves the performance over SDL as long as higher-order interaction terms are incorporated. Even if the DGP is linear (i.e., Panel D), the performance of DeDL is on par with that of SDL, suggesting that debiasing with Neyman orthogonality is fairly robust under model misspecification.

Table A9 Model Misspecification Result with Explicit Individual-Level Higher Order Terms

Panel A: Comparison of Different Estimators Under 1st, 2nd, 3rd, 4th - Order Terms			
Estimator	MAPE	MSE	MAE
LA	34.06% (33.77%, 34.35 %)	18.732 (18.426, 19.038)	2.582 (2.559, 2.604)
LR	29.30% (28.98%, 29.63 %)	12.225 (11.967, 12.482)	1.883 (1.860, 1.907)
SDL	21.78% (21.41%, 22.14 %)	11.453 (11.168, 11.739)	1.671 (1.646, 1.696)
DeDL	15.15% (14.80%, 15.50%)	8.013 (7.706, 8.321)	1.310 (1.277, 1.343)
Panel B: Comparison of Different Estimators Under 1st, 2nd, 3rd - Order Terms			
Estimator	MAPE	MSE	MAE
LA	33.65% (33.35%, 33.96 %)	16.560 (16.297, 16.824)	2.493 (2.472, 2.514)
LR	28.81% (28.49%, 29.14 %)	10.437 (10.246, 10.629)	1.802 (1.784, 1.821)
SDL	21.51% (21.17%, 21.85 %)	9.653 (9.437, 9.870)	1.591 (1.571, 1.611)
DeDL	14.88% (14.56%, 15.19%)	6.476 (6.216, 6.736)	1.230 (1.200, 1.260)
Panel C: Comparison of Different Estimators Under 1st, 2nd - Order Terms			
Estimator	MAPE	MSE	MAE
LA	31.04% (30.74%, 31.33 %)	7.659 (7.485, 7.833)	1.874 (1.853, 1.896)
LR	25.09% (24.80%, 25.39 %)	3.466 (3.362, 3.570)	1.166 (1.146, 1.185)
SDL	17.32% (16.99%, 17.64 %)	2.903 (2.786, 3.019)	0.948 (0.927, 0.968)
DeDL	10.07% (9.74%, 10.40%)	1.328 (1.223, 1.433)	0.591 (0.565, 0.617)
Panel D: Comparison of Different Estimators Under 1st - Order Terms			
Estimator	MAPE	MSE	MAE
LA	1.56% (1.45%, 1.67 %)	0.002 (0.002, 0.003)	0.036 (0.034, 0.039)
LR	0.75% (0.68%, 0.82 %)	0.001 (0.001, 0.001)	0.018 (0.017, 0.020)
SDL	6.32% (6.08%, 6.55 %)	0.157 (0.145, 0.169)	0.215 (0.206, 0.224)
DeDL	6.67% (6.41%, 6.92%)	0.220 (0.210, 0.231)	0.212 (0.195, 0.230)

Notes: All simulations are replicated 200 times, with 95% CIs reported in parentheses. 95% CIs of average absolute treatment effects are (5.03, 5.10), (4.97, 5.04), (4.34, 4.41), and (2.48, 2.55) respectively in Panel A, Panel B, Panel C, and Panel D.

D.5. Imbalanced Covariates

This set of simulations investigates a practical setting where the observed \mathbf{X} distribution deviates from the population. We call this setting *imbalanced covariates*, i.e., \mathbf{X} covariates are imbalanced between any treatment level $\mathbf{t} \in \{0,1\}^m$ and population \mathbf{X} distribution. The imbalanced covariates affect both training and inference stages, invalidating the DeDL estimators. Below, we discuss how to rebalance the covariates to get precise and trustworthy ATE estimates.

We adopt exactly the same simulation setup to Appendix D.3 except that the observed covariates follow a different distribution. Specifically, the last dimension of \mathbf{x} follows the exponential distribution with rate $\lambda \in \{2.0, 1.0, 0.5\}$ instead of the uniform distribution $\mathcal{U}(0,1)$. The ground truth ATEs are still calculated using the uniform distributed \mathbf{x} . Although this setup is relatively simple because the observed covariates uniformly deviate from the true distribution across all treatment combinations while in practice the observed covariates \mathbf{x} may even follow different distributions under different \mathbf{t} , this simulation result still demonstrates the importance of rebalancing covariates.

To reconcile the imbalanced covariates, one can do stratified sampling on sampled units to match the covariate distribution over the population, as we implement in our empirical study. There are also many re-randomization techniques to improve covariate balance in experiments. We refer interested readers to [Morgan and Rubin \(2012\)](#) and [Li et al. \(2020\)](#).

In this simulation study, to keep the discussion simple, we use the same stratified sampling procedure in our empirical study. Specifically, we focus on the imbalanced covariate dimension with exponential distribution. First, we do stratified sampling to keep only the data with \mathbf{x} in the support $[0, 1]^{d_{\mathbf{x}}}$. Then we do stratified sampling to make sure the imbalanced dimension of \mathbf{x} is rebalanced in the sense that the sample sizes in $[0, 1]^{d_{\mathbf{x}}-1} \times [0, 0.5]$ and $[0, 1]^{d_{\mathbf{x}}-1} \times [0.5, 1]$ are the same. Through stratified sampling, we may discard some samples but sacrifice efficiency. We also conducted the stratified sampling with higher accuracy, e.g., the numbers of samples in five buckets with 0.2 bandwidth are the same.

We report the MAEs for different estimators with balanced and imbalanced covariates in Figure A6(a) and (b), respectively. Using the imbalanced covariates for both training and inference, Figure A6(a) reports the comparison results. (We present the complete results in Table A10.) After the stratified sampling, we use the rebalanced covariates for both training and inference; we show the result in Figure A6(b). $\lambda \in \{2.0, 1.0, 0.5\}$ indicates the growing imbalance level. The key observation is that when the data is too imbalanced with $\lambda \in \{1.0, 0.5\}$, DeDL is not precise. After the stratified sampling, debiasing is still trustworthy, reducing MAE compared to SDL.

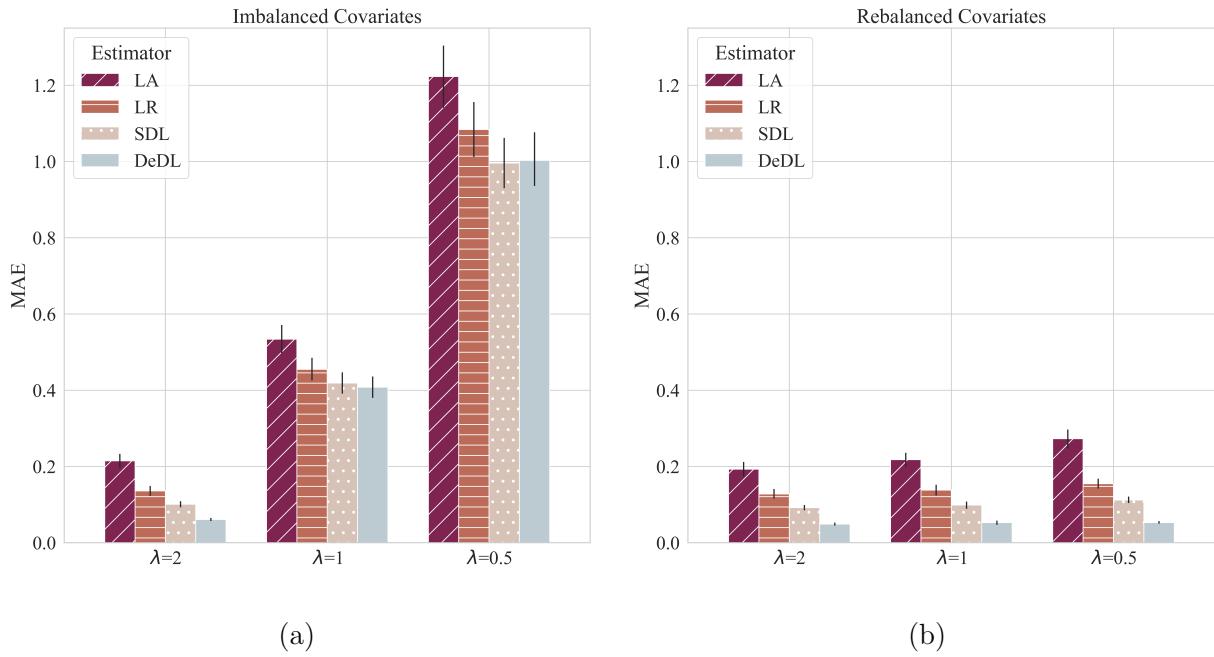


Figure A6 MAE Comparison Among Estimators Under the Imbalanced Covariates Setting. Panel (a) shows the performance of LA, LR, SDL, and DeDL before rebalancing. Panel (b) presents the performance after rebalancing.

Table A10 Imbalanced Covariates Result

Estimator	CRD	MAPE	MSE	MAE	BTI
Panel A: Comparison of Different Estimators Under $\lambda = 2.0$					
Before Covariates Rebalancing					
LA	91.22% (89.99%, 92.45%)	32.03% (28.21%, 35.85%)	0.104 (0.081, 0.127)	0.215 (0.197, 0.233)	87.5% (82.9%, 92.1%)
LR	96.00% (95.32%, 96.68%)	18.92% (16.87%, 20.97%)	0.047 (0.036, 0.059)	0.136 (0.123, 0.149)	91.0% (87.0%, 95.0%)
SDL	96.12% (95.42%, 96.83%)	15.84% (14.09%, 17.60%)	0.024 (0.018, 0.031)	0.101 (0.093, 0.109)	95.0% (92.0%, 98.0%)
DeDL	96.44% (95.74%, 97.13%)	11.10% (9.71%, 12.48%)	0.007 (0.006, 0.008)	0.061 (0.057, 0.065)	95.0% (92.0%, 98.0%)
After Covariates Rebalancing					
LA	93.94% (92.85%, 95.02%)	27.24% (23.10%, 31.38%)	0.102 (0.074, 0.129)	0.193 (0.173, 0.212)	92.0% (88.2%, 95.8%)
LR	96.44% (95.75%, 97.13%)	18.36% (15.69%, 21.03%)	0.044 (0.033, 0.056)	0.128 (0.116, 0.141)	93.5% (90.1%, 96.9%)
SDL	96.62% (95.95%, 97.30%)	14.47% (12.66%, 16.28%)	0.019 (0.015, 0.023)	0.092 (0.085, 0.099)	96.0% (93.3%, 98.7%)
DeDL	97.66% (97.11%, 98.20%)	7.86% (6.66%, 9.05%)	0.006 (0.004, 0.007)	0.049 (0.045, 0.053)	98.0% (96.0%, 100.0%)
Panel B: Comparison of Different Estimators Under $\lambda = 1.0$					
Before Covariates Rebalancing					
LA	87.84% (86.47%, 89.21%)	82.23% (73.27%, 91.18%)	0.519 (0.445, 0.593)	0.534 (0.497, 0.571)	71.5% (65.2%, 77.8%)
LR	90.44% (89.36%, 91.51%)	70.13% (62.66%, 77.60%)	0.371 (0.321, 0.420)	0.455 (0.426, 0.485)	72.0% (65.7%, 78.3%)
SDL	90.53% (89.41%, 91.66%)	71.28% (63.36%, 79.20%)	0.304 (0.260, 0.348)	0.419 (0.391, 0.447)	76.0% (70.0%, 82.0%)
DeDL	90.59% (89.51%, 91.68%)	68.73% (61.16%, 76.30%)	0.291 (0.248, 0.334)	0.408 (0.380, 0.436)	75.5% (69.5%, 81.5%)
After Covariates Rebalancing					
LA	91.44% (90.29%, 92.58%)	32.66% (28.67%, 36.66%)	0.109 (0.086, 0.133)	0.218 (0.199, 0.236)	88.0% (83.5%, 92.5%)
LR	95.84% (95.17%, 96.52%)	20.03% (16.01%, 24.04%)	0.056 (0.036, 0.076)	0.138 (0.124, 0.152)	94.5% (91.3%, 97.7%)
SDL	96.00% (95.23%, 96.77%)	15.04% (12.57%, 17.51%)	0.027 (0.018, 0.036)	0.099 (0.089, 0.108)	96.0% (93.3%, 98.7%)
DeDL	96.47% (95.84%, 97.10%)	9.16% (7.23%, 11.08%)	0.008 (0.004, 0.012)	0.053 (0.047, 0.058)	96.0% (93.3%, 98.7%)
Panel C: Comparison of Different Estimators Under $\lambda = 0.5$					
Before Covariates Rebalancing					
LA	78.41% (76.31%, 80.51%)	186.87% (167.63%, 206.10%)	2.644 (2.289, 2.999)	1.223 (1.141, 1.304)	43.5% (36.6%, 50.4%)
LR	80.84% (78.92%, 82.77%)	167.21% (150.48%, 183.94%)	2.097 (1.811, 2.383)	1.084 (1.012, 1.156)	47.0% (40.0%, 54.0%)
SDL	81.50% (79.51%, 83.49%)	162.02% (144.89%, 179.15%)	1.737 (1.503, 1.972)	0.996 (0.930, 1.062)	51.0% (44.0%, 58.0%)
DeDL	81.34% (79.36%, 83.33%)	163.63% (146.21%, 181.06%)	1.767 (1.528, 2.005)	1.003 (0.936, 1.070)	51.5% (44.5%, 58.5%)
After Covariates Rebalancing					
LA	88.78% (87.40%, 90.17%)	42.56% (37.80%, 47.32%)	0.158 (0.126, 0.191)	0.273 (0.248, 0.297)	84.0% (78.9%, 89.1%)
LR	94.78% (93.85%, 95.72%)	22.60% (20.11%, 25.09%)	0.055 (0.045, 0.065)	0.155 (0.142, 0.168)	88.5% (84.0%, 93.0%)
SDL	94.41% (93.49%, 95.32%)	19.69% (17.41%, 21.97%)	0.026 (0.022, 0.030)	0.112 (0.104, 0.121)	92.0% (88.2%, 95.8%)
DeDL	94.75% (93.87%, 95.63%)	9.27% (8.32%, 10.21%)	0.005 (0.005, 0.006)	0.053 (0.050, 0.057)	93.5% (90.1%, 96.9%)